THE UNIVERSITY *of* EDINBURGH

The role of goal-orientated attention and expectations in visual processing and perception

Matthew Chalk

Doctor of Philosophy
Institute for Adaptive and Neural Computation
School of Informatics
University of Edinburgh
2012

Abstract

Visual processing is not fixed, but changes dynamically depending on the spatiotemporal context of the presented stimulus, and the behavioural task being performed. In this thesis, I describe theoretical and experimental work that was conducted to investigate how and why visual perception and neural responses are altered by the behavioural and statistical context of presented stimuli.

The process by which stimulus expectations are acquired and then shape our sensory experiences is not well understood. To investigate this, I conducted a psychophysics experiment where participants were asked to estimate the direction of motion of presented stimuli, with some directions presented more frequently than others. I found that participants quickly developed expectations for the most frequently presented directions and that this altered their perception of new stimuli, inducing biases in the perceived motion direction as well as visual hallucinations in the absence of a stimulus. These biases were well explained by a model that accounted for their behaviour using a Bayesian strategy, combining a learned prior of the stimulus statistics with their sensory evidence using Bayes' rule.

Altering the behavioural context of presented stimuli results in diverse changes to visual neuron responses, including alterations in receptive field structure and firing rates. While these changes are often thought to reflect optimization towards the behavioural task, what exactly is being optimized and why different tasks produce such varying effects is unknown. To account for the effects of a behavioural task on visual neuron responses, I extend previous Bayesian models of visual processing, hypothesizing that the brain learns an internal model that predicts how both the sensory input and the reward received for performing different actions are determined by a common set of explanatory causes. Short-term changes in visual neural responses would thus reflect optimization of this internal model to deal with changes in the sensory environment (stimulus statistics) and behavioural demands (reward statistics), respectively. This framework is used to predict a range of experimentally observed effects of goal-orientated attention on visual neuron responses.

Together, these studies provide new insight into how and why sensory processing adapts in response to changes in the environment. The experimental results support the idea of a very plastic visual system, in which prior knowledge is rapidly acquired and used to shape perception. The theoretical work extends previous Bayesian models of sensory processing, to understand how visual neural responses are altered by the behavioural context of presetned stimuli. Finally, these studies provide a unified description of 'expectations' and 'goal-orientated attention', as corresponding to continuous adaptation of an internal generative model of the world to account for newly received contextual information.

Acknowledgements

I would like to thank Peggy Seriès for providing support and direction throughout my PhD, and being patient with my stochastic random walk towards completion. I thank Aaron Seitz and Iain Murray for help and direction with the psychophyiscs and modeling work respectively. I thank Alex Thiele for continued supervision, and for arousing my initial interest in studying visual attention. I thank Sophie Denève for early modeling inspiration, and Peter Dayan & Odelia Schwartz for valuable comments on the modeling work. Thanks to David Reichert for interesting discussions, which left me doubting everything. Thanks to my parents and brother for coping while I expounded the details of my research to them (but only on the condition that they read the thesis!). And finally, thanks to my examiners, Barbara Webb & Peter Dayan, for their in depth reading, and insightful criticism of the thesis.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Matthew Chalk*)

Contents

List of Figures

Chapter 1

Introduction

The sensory signals that we receive at each moment in time depend on our environment. In a forest, we will see more green trees than grey buildings; in low light we will not receive much information about colour at all. Contextual information about our environment also influences how incoming sensory signals should best be interpreted. A rectangular object is more likely to correspond to a book if we are in a library, and a brick if we are in a construction site.

There is a growing body of evidence that our sensory system takes such contextual information into account in order to shape and constrain our perception of the world. Contextual changes to visual perception are given a number of different cognitive labels (e.g. 'adaptation', 'expectations or 'learning'), depending on their effect, and the timescale over which they occur. In this thesis, I define *expectation-dependent* changes to sensory processing, as changes to perception and/or neural responses that depend on acquired contextual information about the organism's sensory environment (i.e. the stimulus statistics). In chapter 3, I describe a psychophysics experiment that was conducted to investigate how implicitly learned expectations alter the perceived appearance of simple visual stimulus features. The perceptual changes observed in this experiment can be accounted for by assuming that participants followed a Bayesian strategy, combining their received sensory signals with their learned expectations in a probabilistically optimal manner.

Arguably, the ultimate goal of sensory processing is not to infer the identity or location of objects or features, but to allow us to interact with our environment and perform actions that will lead to a delayed or immediate reward. Thus, under the assumption that the visual system is unable to process information about all aspects of a visual scene (although see section 2.1 for further discussion of this claim), it makes sense to prioritize image features or locations that are relevant in determining which action to perform, at the expense of neglecting information about behaviourally irrelevant image features or locations. Consequently, we would expect visual processing to be modulated by the *behavioural* as well as the *statistical* context of presented visual stimuli. Supporting this claim, is a huge experimental literature on sensory attention,

showing how visual perception and neural responses are modulated by the behavioural goals of the observer.

In this thesis, I use a Bayesian modelling framework (see section 1.2) to investigate why top-down goal-orientated attention alters visual neuron responses as it does. While the word attention is used in the literature to describe many different neural and perceptual changes, in this thesis I investigate *goal-orientated attention,* defined as changes to visual perception and neural responses that depend on the observer's task or behavioural objectives. Thus, I make a distinction between *expectations*, that depend on the subject's belief about the presented stimulus statistics (i.e. how 'likely' different stimuli are), and *goal-orientated attention*, that depends on how stimuli are used to decide which action to perform (i.e. how 'behaviourally relevant' different stimuli are). While previous Bayesian models of visual processing are able to account for the former case, it is not obvious why visual processing should be altered by behavioural demands when the stimulus statistics are unchanged. In chapter 4 I show how a Bayesian framework for modelling visual processing can be extended to account for the effects of behavioural context on sensory neural responses. I use this framework to construct a simple model of visual processing that is able to replicate a number of attention-dependent changes to the responses neurons in the mid-level visual cortices. I show that this model is consistent with, and provides a normative explanation for previous phenomenological models of attention.

1.1 Models of neural processing and cognition

Theoretical models can be used for many different purposes in computational neuroscience. The type of model that is used must be chosen carefully depending on the scientific question that is being addressed. Important choices that must be considered include the level of abstraction, the data that is used to construct the model, and how the model it to be validated or falsified. In this thesis we use two very different types of model to ask questions about the role of behavioural and stimulus context in visual processing and perception.

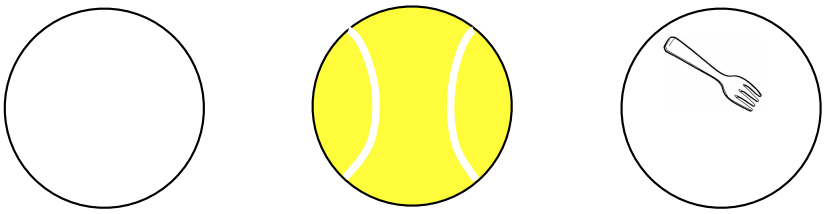
In chapter 3 I use computational modelling approach to understand subjects' behaviour in a psychophysics task. The goal here is both to understand the strategy that subjects took in performing the task, and to rule out 'trivial' explanations for our data. To this end, I constructed several high-level models of subjects' behaviour, which each model representing a different hypothesis about how subject's incorporate learned information about the presented stimulus distribution with incoming sensory information to estimate the presented stimulus feature. I compare these different models using standard statistical tests (such as the Bayesian information criterion), which implicitly assume a trade-off between how well each model fits the data, and the number of model parameters, under the assumption that, if two models fit the data equally well, the model with fewer parameters is to be preferred.

While the computational models proposed in chapter 3 were specifically designed to capture subject's behaviour in our psychophysics task, they also make predictions how people's expectations should influence their perception under a broader range of experimental conditions. Therefore, in considering whether the proposed models provide an appropriate description of subject's behaviour, it was important to also consider how well these models are able to explain known experimental results. In chapter 5 I discuss how the Bayesian model that we propose to explain subject's behaviour in our task could be extended to account for other forms of perceptual bias reported in the experimental literature, and in particular, whether such a model could account for the repulsive perceptual biases that occur following with visual adaptation to a strong motion stimulus.

In chapter 4 I use a normative modelling approach to investigate why goal-orientated attention alters visual neural responses as it does. Normative models rely on the assumption that the brain is well adapted, through evolution and development, to solve the computational problems that it is faced with. The hope is that, by first trying to find good or 'optimal' solutions to an underlying computational problem, we may gain some insight into the computations that actually take place in the brain, and ultimately obtain a functional explanation for experimentally observed neural behaviour.

Care must be taken when using normative models to make predictions about neural activity. In addition to finding an 'optimal' solution to a particular computational problem, the researcher must also make assumptions about how the problem is solved in the brain. For example, the researcher must make assumptions about how the information required to solve a particular problem is computed and encoded in the activity of neural populations. As a result, it may be difficult to make falsifiable predictions about neural activity: if the predictions of the model do not match the data, then this could be because the initial hypothesis about the 'goal' of the system was incorrect, because assumptions about how the system solves the computational problem were incorrect, or even because the underlying normative assumption, that the system behaves in a near-optimal way, was incorrect. However, while it may be difficult to construct a fully falsifiable normative model of neural processing, these models may nonetheless be useful for formalizing and comparing and testing the implications of intuitive ideas about *why* neural systems behave the way they do. Indeed, a normative modelling framework is particularly useful when it allows us to explain a large range of results within a single explanatory framework.

In common with the work described in chapter 3, in chapter 4 I use a normative Bayesian framework (see next section) to model visual processing. However, in contrast to the work described in chapter 3, the main goal in chapter 4 was not to account for the results of a particular experiment, but rather to show in principle how a Bayesian framework can be used to model the effects of behavioural demands on sensory neural responses. In order to demonstrate



Which is furthest away?

Figure 1.1: Prior information about the 'true' size of each object must be combined with available sensory information (the apparent size of each circle) in order to estimate their distance. As the leftmost circle does not represent a real object in the world, we cannot know how far away it is.

how our modelling framework can be used *in practice* to make predictions about attention-dependent changes to neural activity, we needed to make a number of assumptions about how Bayesian inference is performed in the brain. While not all of our assumptions may hold true in reality, our work provides an illustrative example of how Bayesian models of visual processing can be extended to make a number of predictions about the effects of task-dependent attention on visual neuron responses, hopefully opening the door for future work in this area. Interestingly, the results of our simulations are very similar to several existing models of visual attention(Reynolds and Heeger, 2009; Lee and Maunsell, 2009; Ghose, 2009), allowing us to make a link between functional and phenomenological levels of description of visual attention.

1.2 Bayesian models of perceptual processing

In this thesis I consider a *Bayesian* description of visual processing (Neisser, 1970; Gregory, 1970; Lee and Mumford, 2003; Knill and Pouget, 2004), in which the assumed 'goal' of the visual system is to infer the true state of the world from received sensory signals. Because of the inherent ambiguity of sensory signals (figure 1.1), knowledge about the world is expressed in the form of a probability distribution: the organism combines their available sensory information with prior knowledge about the world in order to evaluate the posterior probability distribution over possible world-states. In the following sections, I describe this Bayesian modeling framework in more detail, before outlining the open questions to be addressed in this thesis.

1.2.1 Introduction to Bayesian inference and decision making

Bayesian theory describes explicitly how ambiguous sensory information should be combined with prior knowledge about the world in order to make optimal perceptual decisions (Jaynes, 1986; MacKay, 2003). The *posterior* probability associated with different states of the world

(p(world-state|sensory signal)) is computed by combining *prior* beliefs about the world (p(world-state)) with a *likelihood* model that describes how sensory signals are generated (p(sensory signal|world-state)), according to Bayes' rule:

$$p(\text{world-state}|\text{sensory signal}) = \frac{p(\text{sensory signal}|\text{world-state})\, p(\text{world-state})}{p(\text{sensory signal})}. \tag{1.1}$$

The denominator in this expression represents a normalization constant, which ensures that the posterior probability over all different possible world-states sums to one.

In order to use the posterior probability distribution to make optimal decisions, or perceptual judgments, we need to specify a function (U(decision; world-state)) that quantifies the utility associated with making different decisions, given a particular state of the world (Yuille and Bulthoff, 1996; Körding and Wolpert, 2006). A decision is then made to maximize the expected utility, averaged over the posterior distribution over world-states:

$$\text{decision} = \arg\max_{\text{decision}} \langle U(\text{decision}; \text{world-state}) \rangle_{p(\text{world-state}|\text{sensory signal})}. \tag{1.2}$$

Thus, to formulate how perceptual decisions can be made optimally from ambiguous sensory information, we must specify three ingredients (Simoncelli, 2009):

- The prior probability for different states of the world: p(world-state).
- The likelihood function, describing how sensory signals are generated: p(sensory signal|world-state).
- A utility function, describing the value associated with different perceptual decisions, given the state of the world: U(decision; world-state).

To see how each of these ingredients impact on decision making, consider a doctor tasked with diagnosing a patient. In this case, possible diseases correspond to the 'world-state', while observed symptoms correspond to the 'sensory signal'. The probability that a patient has a particular disease will depend on both the likelihood that the disease produces the observed symptoms, and how rare the disease is. For example, there may be a high probability that meningitis causes an increase in temperature (i.e. p(high temp|meningitis) is large), but meningitis is very rare (i.e. p(meningitis) is small) the posterior probability that it is responsible for the observed symptoms will also be small (as p(meningitis|high temp) $\propto p$(high temp|meningitis) p(meningitis)). However, the doctor's diagnosis will also be influenced by the cost associated with making different errors (U(diagnosis; disease)): wrongly diagnosing the patient as having meningitis may not be a problem, compared to the cost of wrongly diagnosing them as being well. Thus, the utility function may bias the doctor towards diagnosing the patient as having meningitis, despite the small posterior probability associated with this eventuality.

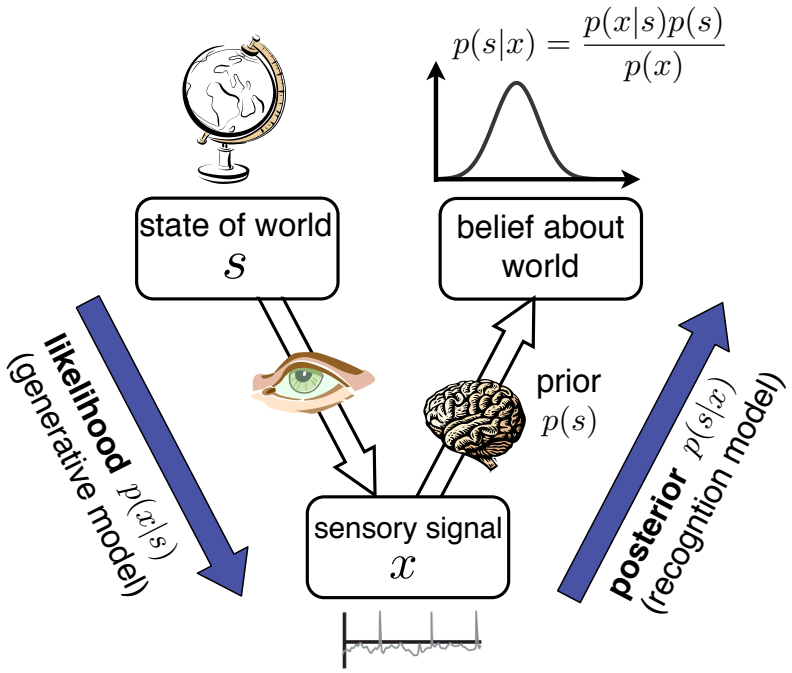


Figure 1.2: Schematic illustration of the Bayesian view of perception (adapted from (Whiteley, 2008)). A generative model (*likelihood*; $p(x|s)$) describes how the state of the world (s) produces the sensory signal (x). The brain is hypothesized to learn a recognition model (*posterior*; $p(s|x)$), which takes existing beliefs (*priors;* $p(s)$) into account in order to make inferences about the state of the world.

1.2.2 Bayesian inference in the brain

In the previous section we described how, in theory, optimal decisions can be made from ambiguous sensory information. The *Bayesian brain hypothesis* ('BBH') postulates that decision making in the brain operates according to these same principles. Thus, sensory processing is hypothesized to correspond to a process of unconscious inference, where incoming sensory signals are used to infer the posterior probability associated with different states of the world, according to Bayes' rule (equation 1.1) (Knill and Pouget, 2004; Knill and Richards, 1996a). This information is then propagated to higher areas of the brain, where decisions are made in order to maximize the expected utility (equation 1.2) (Yuille and Bulthoff, 1996; Körding and Wolpert, 2006; Platt and Glimcher, 1999).

Figure 1.2 illustrates the Bayesian view of sensory processing. Objects in the world (s) generate the received sensory signals (x) with probability, $p(x|s)$. The hypothesized goal of sensory processing is to invert this model, inferring the posterior distribution of world states, given the received sensory signal ($p(s|x)$). To do this, the brain is assumed to learn an internal model describing how sensory signals are generated, which is combined with prior beliefs about the world ($p(s)$) according to equation 1.1.

The BBH makes a number of predictions about how we should perceive the world. First, it implies that we learn an internal model of the world, with prior beliefs that reflect the statistics of the sensory signals that we experience. These prior beliefs should be combined probabilistically with our received sensory signals according to Bayes' rule: the more ambiguous or noisy sensory signals are, the more strongly prior knowledge about the world should influence what we perceive (Stocker and Simoncelli, 2006a; Weiss et al., 2002; Girshick et al., 2011). Second, different sources of sensory information should be combined probabilistically, with their impact on perception depending on how reliable they are. For example, in low light, we should rely more on our sense of hearing than on our sight (Ernst and Banks, 2002; Battaglia et al., 2010). A number of psychophysics experiments have been conducted which support these behavioural predictions (discussed in section 2.4).

In addition to investigating the behavioural predictions of the BBH, researchers have also tried to understand how Bayesian inference might be implemented in the brain. One approach has been to investigate how probability distributions are encoded via neural population responses (Ma et al., 2006; Fiser et al., 2010; Deneve, 2008a; Rao, 2004) (discussed in section 5.3). Another approach has been to try and 'derive' the response properties of sensory neurons from first principles, based on the statistics of natural images (Hyvärinen, 2010; Simoncelli and Olshausen, 2001; Karklin and Lewicki, 2009; Olshausen and Field, 1996) (discussed in section 5.1). The assumption here is that the brain learns an internal model describing how sensory signals are generated by a limited number of primitive image features. Thus, the features encoded by neurons in the early visual cortex can be learned directly from the statistics of the sensory signals themselves.

1.2.3 Overview of thesis

In this thesis I examine how and why changes to behavioural and stimulus context alter visual processing and perception. In chapter 2, I review existing approaches to modelling sensory attention and expectations. In chapter 3, I provide experimental evidence indicating that visual perception is highly adaptable, such that the perceptual appearance of simple visual features varies constantly depending on what we expect to see. I use a Bayesian modeling framework to investigate the functional role of these perceptual changes (why do they occur?). In chapter 4, I investigate how Bayesian models of visual processing can be extended to account for the effects of behavioural demands; providing a functional explanation for observed attention-dependent changes to visual neuron responses. Finally, in chapter 5 I discuss the implications of my experimental and theoretical work, its limitations, and the scope for using Bayesian models in the future to investigate contextual changes to visual processing and perception.

Chapter 2

Theories of goal-orientated attention and expectations

In 1890, William James began his famous description of attention with the claim, "*Everyone knows what attention is*" (James, 1890). This is hard to contest: from our earliest schooldays we are told to 'pay attention' to our teacher, under the assumption that this will help us to hear more clearly and retain in our memory what they have to say. The same can be said of expectations: we all know what it feels like to expect a stimulus (e.g. when we complete an unfinished sentence), or to have our expectations violated (Summerfield and Egner, 2009). However, despite (or perhaps because of) this everyday understanding, defining precisely what expectations and attention are, and why they are necessary, is not easy. Indeed, just six years after William James' famous quote, Groos wrote, "*To the question 'What is Attention?' there is not only no generally recognized answer, but the different attempts at a solution even diverge in the most disturbing manner*" (Groos, 1896). This is arguably as true today as it was then (Sutherland, 1998; Anderson et al., 2011).

One reason for this confusion is that attention is implicated in a huge range of cognitive phenomena, ranging from perception (Raymond, 2000), to learning and memory (Crist et al., 2001; Desimone, 1996). Attention can be directed towards external stimuli, or internal thoughts (Chun et al., 2011); be controlled voluntarily, based on internal goals, or involuntarily, based on the salience of presented stimuli (Corbetta and Shulman, 2002; Yantis, 2000). As a result, it has been suggested that attention is not unitary, but rather, corresponds to multiple different perceptual and cognitive processes that govern how information is selectively processed in the brain (Chun et al., 2011).

Despite the diverse effects of attention, certain themes are observed throughout the literature. In particular, a general assumption is that the brain receives more information than it is able to deal with, and therefore, must select some sources of information for further processing, while discarding others. This leads to several questions, which we discuss separately in each

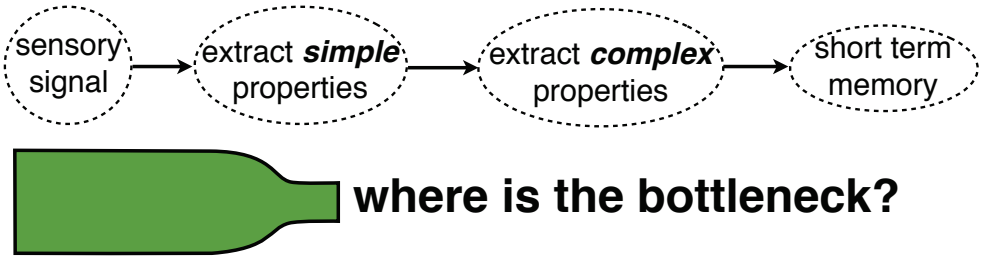


Figure 2.1: Simple schematic of perceptual processing. Arrows denote the direction of feed-forward processing. Early theories of attention assumed that there is a bottleneck which restricts how much information can be propagated through the processing hierarchy. Attention would act as a selective filter, ensuring that only important information was able to reach short term memory and influence behaviour. There has been much debate about the stage of processing that attentional selection occurs (i.e. where the processing 'bottleneck' lies): selection could occur at an early on, very late, or in a graded manner, at multiple stages of processing.

of the following sections. First, where does the bottleneck lie: under what circumstances, and at what stage of processing, is attentional selection necessary? Second, what factors determine which information is selected? Third, how does selective attention give rise to the perceptual and neurophysiological effects that are observed experimentally? A final question, concerning the low-level neural mechanisms underlying attentional selection, is considered beyond the scope of this review.

In this thesis we are interested in how visual processing and perception are influenced by the context of presented stimuli. Indeed, which stimuli are attended to will depend on the context of the organism's current behavioural demands (which stimuli are behaviourally relevant?) and their sensory environment (which stimuli are statistically likely?). In chapter 1 we distinguished between the perceptual effects of behavioural and statistical context, which we labelled as 'goal-orientated attention' and 'expectations' respectively. In section 2.4 we discuss whether this distinction is indeed necessary to describe the neural and perceptual changes that are observed experimentally: do similar cognitive phenomena underlie the perceptual changes that occur in each case?

2.1 Why attend?

Why is attention necessary? Initially, the answer seems intuitively obvious: in a crowded restaurant, we need to attend to the person speaking to us in order to hear them clearly, while filtering out distractions from other nearby voices (the so-called 'cocktail-party' problem (Cherry, 1953)). On further reflection however, it becomes apparent that this intuitive explanation makes strong assumptions about the *limited resources* that are available to process sensory informa-

tion. If there were no limited resources it would be possible to fully process all incoming sensory information, without needing to filter out irrelevant distractions (although see later for a discussion of how selective attention could be in the absence of limited resource constraints).

Early attempts to define how perceptual processing is limited, and thus why attention is necessary, were strongly influenced by the newly emerging mathematical theory of information (Shannon, 1948; MacKay, 2003). A central idea in information theory is that communication channels have a *limited capacity*: a fundamental limit to the rate at which information can be transferred. The concept of a limited capacity channel provided the basis for Broadbent's 'filter theory' (Broadbent, 1958), which was one of the first attempts to describe formally why attentional selection is necessary. Broadbent hypothesized that perceptual processing is divided into two qualitatively distinct stages: an initial stage, where 'simple' properties of sensory signals (e.g. pitch and location of sounds) are extracted in parallel, and a second stage, where more complex properties (e.g. the identity or meaning of spoken words) are extracted (figure 2.1). Crucially, the second stage of processing is assumed to have a very limited capacity, so that it is unable to process information about multiple stimuli simultaneously. Attention deals with this limited capacity constraint by selectively filtering incoming sensory information, so that only information about attended stimuli reaches high-level processing, while information about unattended stimuli is discarded.

Broadbent's filter theory involves *early selection*, with only attended stimuli undergoing high-level processing. Therefore, experimental results indicating that, contrary to Broadbent's theory, subjects are sometimes able to perceive high-level features of unattended stimuli (Corteen and Dunn, 1974), led some researchers to propose an alternative theory, that attentional selection occurs at a late stage of perceptual processing (*late selection*) (Deutsch and Deutsch, 1963; Duncan, 1980). According to these researchers, all sensory signals undergo high-level processing, regardless of whether they are attended or not. Attention then acts to select which information reaches short term memory and therefore, is able to influence responses. Indeed, the question of whether selection occurs at an early or at a late stage of perceptual processing was debated for many years (see (Driver, 2001) for a historical review), with experimental evidence found to support both views (Eriksen and Eriksen, 1974; Francolini and Egeth, 1980). To account for seemingly contradictory experimental results on both sides of this debate, Treisman suggested an intermediate possibility, where selection occurs in a graded manner, at multiple levels of processing (*attenuation*) (Treisman, 1960, 1969).

More recently Gottlieb et al. proposed that attention could be viewed as a form of cognitive decision: in the same way that a motor decision requires selecting one of many many possible actions, an 'attentional-decision' would involve selecting only the most important sensory information for further processing (Gottlieb and Balan, 2010). This idea is closely related to earlier limited capacity theories of attention: it assumes that only a subset of incoming sensory

information can be processed effectively, and that the role of selective attention is to determine which sensory information will be passed on for high-level processing.

While attractive in their simplicity, a potential criticism of these theories is that the concepts that they rely on – *limited capacity* and *perceptual load* – are extremely difficult to quantify. Further, these concepts are unlikely to be constrained by information theory alone: information theory describes 'how much' information is present in sensory signals, but does not say anything about their semantic content (e.g. their behavioural relevance, or how to extract high-level features). Thus, a full understanding of attentional selection requires a richer description of perceptual processing, explaining how useful high-level representations are computed from incoming sensory signals.

One way to make the 'limited resource' constraint more concrete is to set out explicitly the *computational problem* faced by the visual system. Thus, the necessity for attention is dictated by the complexity of the problem that must be solved. Tsotsos et al. posited that the general 'vision problem' can be formulated as an unbounded visual search (i.e. locating objects within a scene). They showed that the problem is computationally intractable in general, such that the time taken to locate a given object or feature scales very quickly with the size of the image (Tsotsos, 1989). Tsotsos et al. proposed that, by limiting the search to selected regions of interest, visual attention reduces the complexity of the problem, facilitating the construction of a tractable algorithm that approximates the general search problem (Tsotsos, 1990; Tsotsos et al., 1995). A potential criticism of this approach is that, complexity theory just provides a way of quantifying how the number of steps required to solve a problem scales with its size. However, the visual system only has to solve problems of a maximum fixed size, and thus, how the problem scales may not be important.

Another approach, that avoids making explicit assumptions about the limited available resources, is to consider how the visual system extracts high-level representations from incoming sensory signals. In this approach, attentional selection emerges as part of an algorithm to compute high-level representations of a particular desired form. For example it has been proposed that, by dynamically selecting which sensory information is propagated to higher cortical areas, attention could provide a mechanism for computing high-level representations that are invariant to the position and scale of objects in the visual world (Hudson et al., 1997; Anderson and Van Essen, 1987; Olshausen et al., 1993; Deco, 2004). An alternative idea, that has been highly influential in models of attention, is that selective attention is required to compute high-level representations in which low-level features (e.g. colour, location) are bound together into coherent unitary percepts (Treisman, 1960, 1969; Reynolds and Desimone, 1999).

There is a large degree of overlap between models which view the role of attention as shaping the high-level representation, and models in which attention is required to find the solutions to a particular computational problem: a high-level representation that is invariant

to the size and location of visual objects, or in which low-level features are bound together into a coherent whole, will enable many of the computational problems faced by the visual system, such as object recognition and visual search, to be performed more easily. Arguably, both classes of model provide a richer description of attention than earlier theories based on the concept of a limited capacity channel. However, they also face the potential criticism of presenting too narrow a view of visual processing; as a mechanism to solve a particular computational problem, or to construct high-level representations of a particular form.

2.1.1 Bayesian models of attention

Most theories of attention assume, either explicitly or implicitly, that the necessity for attentional selection comes from the fact that the brain is unable to fully process all incoming sensory information. In contrast, recent 'Bayesian' models of attention have argued that in many cases, attentional selection represents the optimal strategy, even in the absence of any limited resource constraints (Dayan et al., 2000). As discussed earlier (section 1.2), these models are based on the hypothesis that visual processing corresponds to a process of probabilistic inference, where the hidden state of the world is inferred by combining sensory signals with prior beliefs according to Bayes' rule (equation 1.1). Thus, perceptual inference depends on both the observer's internal model describing how their received sensory signals are generated (the *likelihood*), and the probability that they associate with different states of the world (the *prior*). In order to make optimal inferences about the hidden state of the world, the observer's internal model should be updated to incorporate new information about their environment. This information could be delivered explicitly, through instructions or sensory cues that indicate which stimuli are most likely to be presented (Yu and Dayan, 2005a,b); or it could be inferred directly from the sensory signals themselves, for example, when the observer is presented with novel stimuli that violate the statistics of the natural environment (Yu et al., 2009; Liu et al., 2009). In both cases, attention-dependent changes to perceptual processing would correspond to changes in the observer's internal model that take place in order to account for newly received information about their environment.

Experimentally, it is clear that selective attention can be influenced by the *behavioural relevance* of sensory signals, as well as their statistics (Pestilli and Carrasco, 2005). However, an ideal Bayesian observer who has learned a perfect internal model of their environment should not alter their perceptual inference strategy depending on behavioural demands: they should use Bayes' rule to infer the hidden state of both task-relevant, and task-irrelevant aspects of the world. In other words, if the computational resources available for perceptual processing are unlimited, there is no need to throw away task-irrelevant sensory information. Thus, to understand how attention is shaped by behavioural demands, we need to consider the 'limited resource' that prevents the observer from making optimal inferences about the world. In the

context of the BBH, perceptual processing can be constrained in three different ways: by the structure of the internal model that can be learned, by the limited quantity of available training data (e.g. when the external environment changes very fast), or by the computational power required to compute the posterior probability distribution from Bayes' rule. These factors are closely related: the complexity of the internal model is likely to be constrained by the limited data available to learn model parameters, as well as the feasibility of perceptual inference. The way in which computational resources are limited may also depend how Bayesian inference is implemented in the brain: for example, by the limited number of neurons available to encode the posterior distribution, the high metabolic cost of firing a neural spike (Lennie, 2003), or due to constraints on cortical connectivity.

In one of the first Bayesian models of selective attention, Dayan & Zemel proposed that perceptual processing is constrained by the limited number, and therefore, the necessarily broad tuning of neural receptive fields (RFs) (Dayan and Zemel, 1999). They hypothesized that the broad RFs of visual neurons result in a mismatch between the hidden causes in the internal model, which extend over a large region of space, and the stimuli typically used to study visual attention, which are more spatially localized. They proposed that attention compensates for this mismatch by imposing a 'task-dependent' prior which favours behaviourally-relevant spatial locations. As a result, sensory signals from irrelevant spatial locations are filtered out, so that the features (e.g. the orientation) of stimuli presented at attended locations are estimated more accurately and with greater certainty than the features of stimuli presented at unattended locations. Given certain assumptions about how probability distributions are encoded by neural activity, Dayan & Zemel showed that this reduction in uncertainty could lead to an increase in the activity of neurons tuned to attended locations.

Several Bayesian models of attention have been proposed to account for perceptual performance in the 'Eriksen task' (Yu et al., 2009; Dayan and Solomon, 2010; Dayan, 2008; Liu et al., 2009), where subjects are asked to report the identity of a target letter presented amongst nearby distractors (Eriksen and Eriksen, 1974). As with the work of Dayan & Zemel, these models hypothesize that perceptual performance is limited by the broad size of neural RFs, which integrate sensory signals from both target and distractor stimuli. As a result, when subjects make a rapid response in the task, their performance is strongly degraded by the distractors. However, as more sensory observations are accumulated, these models predict that the subject will learn to disregard information from neurons with broad RFs. Thus, given more time to make their responses, task-irrelevant sensory signals from nearby distractors will be filtered out, so that perceptual performance is not diminished by the distractors. Interestingly, these models do not include an explicit attentional mechanism to filter out irrelevant sensory signals; attentional effects emerge automatically as the observer updates the inferred posterior distribution to account for newly received sensory information.

Interestingly, Bayesian models have been used to account for aspects of attention that appear to be suboptimal. For example, Lavie et al. found that in easy tasks, subjects are often unable to ignore irrelevant stimuli that they are able to ignore when the task is hard. To explain these results, Lavie proposed that the degree to which unattended sensory information is filtered depends on the *perceptual load*: when the task is easy, irrelevant sensory information is always processed (even at the cost of worse performance on the relevant information), while in difficult tasks, no capacity remains, and irrelevant information is effectively removed (Lavie, 2005). An alternative explanation was proposed by Dayan (Dayan, 2008), who showed that a similar effect is predicted for a Bayesian observer which optimally integrated information from neurons of varying receptive field (RF) size. In the 'low load' case, where a task-relevant target stimulus is presented alone, neurons with large RFs are informative about the target, and thus automatically play a role in perceptual inference. Since these neurons are also be influenced by a distractor, there is a reduction in performance when a distractor is presented. In the high-load case, where multiple irrelevant stimuli are presented near to the target, only neurons with small RFs that provide reliable information about the target are used for inference, and the deleterious effect of a distractor is decreased.

In addition to the limited number (and thus, the broad tuning) of neural receptive fields (RFs), there have been various other suggestions as to how the internal model could be constrained. For example, Whiteley et al. proposed that certain dependencies between hidden variables are neglected by the visual system, which represents a factorized approximation of the true posterior. In this view, attentional effects emerge as part of an approximate inference algorithm that selectively improves aspects of the internal model that are most relevant to the organism (Sahani and Whiteley, 2007, 2011; Whiteley, 2008). Other researchers have proposed that the visual system is constrained to learn a simplified internal model, in which the identity and location of visual objects are assumed to be independent (Rao, 2005; Chikkerur et al., 2010). Alternatively, it has been proposed that the visual system learns an internal model in which only one object at a time is explicitly represented (Reichert et al., 2011a; Chikkerur et al., 2010).

In chapter 3 we describe an experiment that was conducted to investigate how visual perception is altered when subjects are presented with novel stimulus statistics. In this case, we would expect changes in visual perception even in the absence of any limited resource constraints, as the internal model is adapted to reflect the presented stimulus statistics (see section 2.4). Indeed, while it is possible that the behavioural task that subjects were asked to perform played a role in producing the perceptual changes that were observed, a purely 'task-independent' Bayesian model, in which subjects' are assumed to learn the presented stimulus statistics (irrespective of their behavioural relevance), is sufficient to explain our results.

In chapter 4 we describe modeling work investigating how visual perception is influenced

by behavioural demands, in the absence of any changes to the presented stimulus statistics. As discussed previously, if people were to learn a perfect internal model of their sensory environment, we would not expect visual processing to be altered by changes to their behavioural demands alone. In our work, we postulate that attentional modulation occurs because the stimuli that are relevant to the task differ from the image features in the agent's internal model. While there are many possible ways that the internal model could differ from their external environment, we postulate a particular form of model mismatch, in which the high-level features in the agent's internal model are more spatially localized than the image features that are relevant to the particular task. As argued by Dayan & Zemel (described above), such a model mismatch could occur due to the limited number of visual neurons, which forces them to have large RFs (Dayan and Zemel, 1999). Alternatively, such a model mismatch could emerge if the visual system tries to learn a simple internal representation for learning new behavioural tasks, in which the reward received for a given action is assumed to depend on a limited number of spatially distributed image features. A final possibility is that broad RFs do not come about as a result of a resource bottleneck at all, but reflect the fact that the image features relevant to most real-world tasks (e.g. objects & faces) are distributed over a broad region of space.

2.2 What should be attended?

Assuming that we understand why attentional selection is necessary, how does the brain know what sensory information to select, and what to discard? Broadly speaking, attention can be controlled in two different ways: *voluntarily*, depending on the task-demands and goals of the observer ("look for the red target"), and *involuntarily*, depending on the properties of the sensory signals themselves (e.g. an intrinsically conspicuous stimulus, such as the red jacket in figure 2.2, will automatically attract attention) (Corbetta and Shulman, 2002).

In our work, we are interested in how visual processing adapts in response to changes in the behavioural or statistical context of presented stimuli. Thus, we focus on voluntary attention, which we view as corresponding to short-term changes in visual processing, that allow the visual system to preferentially process information about stimulus features or locations that are task-relevant or statistically likely. In contrast, involuntary attention, where certain stimulus features are automatically selected regardless of their contextual likelihood or relevance to the task at hand, could emerge due to evolutionary optimization of the visual system towards typically encountered behavioural tasks (see section 5.1 for further discussion).

An experimental paradigm that has been particularly important in understanding top-down attentional control is *visual search*, where people are asked to detect a target stimulus as quickly as possible, from among a varying number of non-target distractors (analogous to figure 2.2b). Visual search experiments provided the main impetus for Treisman's *feature integration the-*

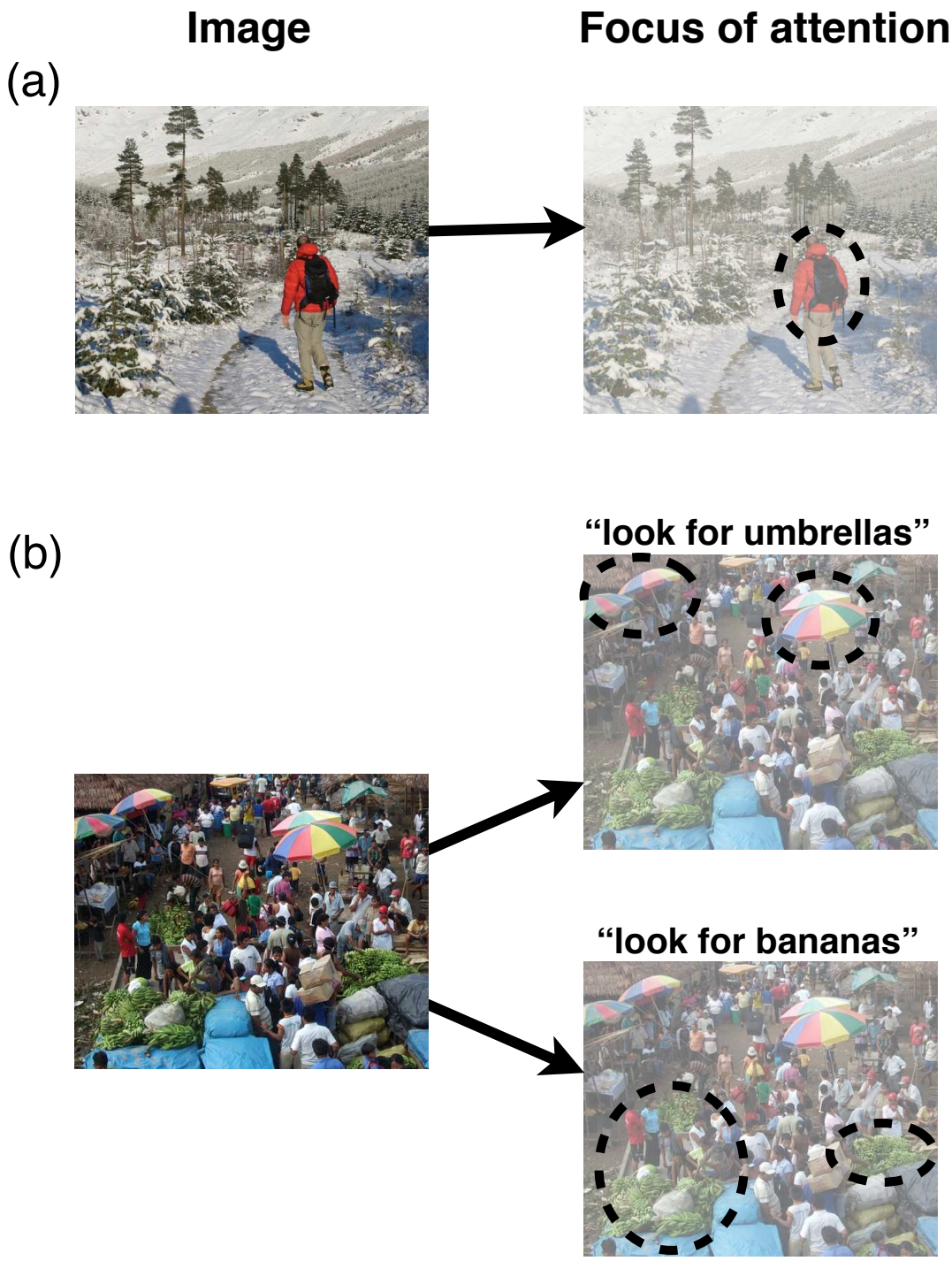


Figure 2.2: Illustration of involuntary and voluntary attentional control. (a) Attention is automatically directed towards the red jacket in the image, as it differs strongly from the (mostly white) background. (b) In this cluttered market scene, attention can be directed towards different locations, depending on the goal of the observer (i.e. whether they are searching for bananas or umbrellas). Visual search experiments, where subjects have to search for a target stimulus presented amongst distractors, provide a simple analogue of this real life scenario.

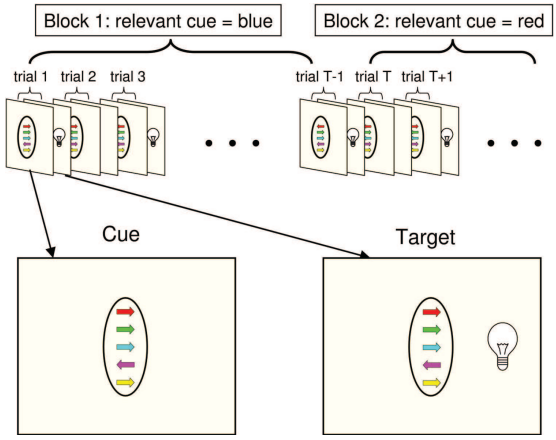


Figure 2.3: Schematic of experimental paradigm to investigate how attentional control is learned from experience (from (Yu and Dayan, 2005b)). On each trial, multiple sensory cues are presented, followed by a target stimulus presented at one of two locations after a variable delay. The subject's task is to detect the target stimulus as quickly as possible. Unknown to the subject, only one of the cues is predictive of the target location. The identity of the task-relevant cue changes in each block of trials. The subject must use their acquired experience in the task to infer which cue is relevant, so that they can direct attention towards the location that the target is most likely to be presented.

ory, in which attention is required to group together simple features (e.g. colour, orientation) into coherent percepts (Treisman and Gelade, 1980; Wolfe et al., 1989; Cave and Wolfe, 1990; Mozer and Baldwin, 2008) (see previous section). Related to the issue of perceptual grouping, there has been much debate about whether attentional control is primarily space-based (Posner et al., 1980), or object-based (Duncan, 1984). Conventionally, people were assumed to perform visual search by attending to each spatial location in turn (i.e. a moving 'spotlight' of attention). However, this idea was challenged by experiments showing that when one aspect of an object is attended (e.g. its shape), other features associated with the object (e.g. colour, motion) are also selected, even when the attended and ignored objects are spatially superimposed (O'Craven et al., 1999; Roelfsema et al., 1998). However, space-based and object-based models of attentional control are not necessarily mutually exclusive – people may use different attentional strategies in different situations. Indeed, how the attentional control strategy depends on the experimental setup (i.e. task and stimulus), and whether similar neural mechanisms underlie both space-based and object-based attentional control are active areas of research (Yantis and Serences, 2003; Hopfinger et al., 2000).

Top-down attention is generally assumed to select behaviourally relevant stimuli for increased perceptual processing. However, in realistic situations, where people are not told what to attend to, they must use their accumulated sensory experience and behavioural feedback to

infer which stimuli are the most relevant and therefore, should be attended. Experimentally, it has been shown that implicitly learned (i.e. subconscious) information about previously observed stimulus contexts can help guide selective attention, resulting in improved behavioural performance (Chun and Jiang, 1998; Chun, 2000; Chun and Nakayama, 2000). Recent theoretical work has proposed that this 'attentional learning' could be explained by assuming that people accumulate sensory information about their environment in a statistically optimal way (Dayan et al., 2000; Yu and Dayan, 2005b; Eckstein et al., 2004; Gershman et al., 2010). For example, Yu & Dayan considered a hypothetical extension to the 'Posner task' (Posner et al., 1980), in which a subject is presented with multiple sensory cues, followed by a target presented at one of two spatial locations (figure 2.3). Unknown to the subject, only one of the sensory cues provides task-relevant information about the spatial location that the target stimulus is most likely to be presented. Further, the identity of the task-relevant cue changes after a random number of trials. Yu & Dayan constructed a normative model to describe how the subject should use their acquired sensory experience to infer which cue is task-relevant. Interestingly, the model differentiated between two kinds of uncertainty: 'expected uncertainty', which comes from the known unreliability of the predictive cue; and 'unexpected uncertainty', due to changes in context (i.e. the identity of the task-relevant cue) that produce unexpected observations. While both types of uncertainty would promote learning about the context, they would interact in a complex manner; for example, if the cue was known to be very unreliable (i.e. high 'expected uncertainty'), changes in context (that introduce 'unexpected uncertainty') would have less impact on learning. Yu & Dayan's work highlights the potential complexity of attentional learning, which will likely depend not only on the subject's knowledge of the stimulus context, but also on the reliability of contextual information, and how much the subject trusts their own knowledge.

We conducted a psychophysics experiment in which subjects were not explicitly told which stimuli were most likely to be presented, but had to learn the stimulus statistics from experience (chapter 3). Later, we simulated a behavioural task in which subjects were required to learn which stimuli were task-relevant from feedback in the task (chapter 4). In both these studies, our main aim was not to understand the learning process itself (although see section 5.2), but rather, how acquired sensory experience and task-feedback is used to control visual attention and expectations. In our psychophysics experiment, we sought to understand how implicitly learned stimulus expectations alter visual perception (see section 2.3.1). The purpose of our modeling work was to construct a theoretical framework that could predict the effects of attention directly from the properties of the behavioural task. This work was motivated by recent experimental evidence indicating that small changes to the behavioural task, leading different sizes of 'attentional focus', can produce qualitatively different changes in neural responses and perception (see section 2.3.2).

2.3 How does attention affect visual processing?

2.3.1 Does attention alter appearance?

A huge degree of experimental effort has been expended on studying how selective attention alters perception. Providing a comprehensive review of such a large body of literature is beyond the scope of this thesis (see (Pashler, 1998; Driver, 2001) for a historical overview). Instead, we focus our discussion on a single area of debate that is relevant to our experimental work (chapter 3): does attention alter the subjective appearance of visual stimuli?

Testing whether attention alters visual appearance is not as simple as it might at first seem. Quantifying the subjective appearance of visual stimuli requires relying on people's self-reports about what they perceive, which could also be influenced by decision biases, memory, or eye movements (Prinzmetal et al., 1997, 1998). Carrasco et al. conducted an experiment to investigate whether attention alters perceived stimulus contrast, which was designed carefully to avoid these potential confounds (Carrasco et al., 2004). Subjects were presented with two gratings, each at a different orientation and contrast, located to either side of a central fixation point. They were asked to "report the orientation of the stimulus that is higher/lower in contrast". This question was designed to reduce bias by placing emphasis on the orientation discrimination task, while 'disguising' the comparative judgement of contrast that was the main interest of the study. On some trials, a dot (the 'cue') appeared briefly at one of the grating sites immediately before the gratings were presented, attracting involuntary attention towards this location. Carrasco et al. found that the point of subjective equality ('PSE'), where subjects are equally likely to report the orientation of either grating, was shifted by the cue; for the two gratings to appear to have the same contrast, the physical contrast of the uncued (i.e. unattended) grating had to be higher than the cued (i.e. attended) grating. This result led Carrasco et al. to conclude that attention increases perceived stimulus contrast. Subsequent work, using a similar experimental design, has reported that involuntary attention can also increase perceived spatial frequency (Gobell and Carrasco, 2005), colour saturation (Fuller and Carrasco, 2006), speed (Turatto et al., 2007) and stimulus size (Anton-Erxleben et al., 2007).

Despite the attempts of Carrasco et al. to rule out alternative 'non-perceptual' explanations for their results, some researchers have disputed whether the biases that they observed were perceptual in origin (Schneider, 2006; Prinzmetal et al., 2008; Schneider and Komlos, 2008). Schneider et al. hypothesized that, rather than altering the perceived stimulus contrast, attention could affect the decision mechanism, causing subjects to report the cued stimulus as having a higher contrast, despite the two stimuli being perceptually identical (Schneider and Komlos, 2008). To test this hypothesis, Schneider et al. changed the type of decision that subjects were asked to perform from a comparative judgement ("which target has higher contrast?") to an equality judgement ("are the two targets equal in contrast") that is resistant to decision bias.

When subjects were asked to make a comparative judgement, Schneider et al. obtained similar results to Carrasco et al., with attention-dependent shifts in the PSE. However, when subjects were asked to perform an equality judgement, these attentional effects disappeared, and there was no shift in the PSE. These results were used by Schneider et al. to support their hypothesis that the reported effects of attention on perceptual appearance can be explained by decision biases, and that attention does not alter appearance.

A shared characteristic of the reported attention-dependent changes in perceptual appearance is that they are consistent with an increase in the saliency of the attended stimulus (e.g. all else being equal, a high contrast stimulus will be more salient than a low contrast stimulus). This fact alone might give rise to concerns of the type raised by Schneider et al.: if the saliency of the attended stimulus is increased, subjects could be biased to report that they perceive features that are consistent with this increase in saliency, despite the perceptual appearance of the stimulus being unchanged. One way to get around this problem would be to investigate how attention alters the perceptual appearance of visual features that do not impact on saliency, such as the motion direction or orientation of an isolated stimulus. Indeed, previous studies have shown that feature-based attention can modulate how different motion stimuli are perceptually combined, thus altering their perceived direction (Chen et al., 2005; Tzvetanov et al., 2006). For example, Chen et al. (Chen et al., 2005) found that attending towards one of two overlapping motion signals reduces the degree of repulsion between them, so that the non-attended motion direction is perceived as being closer to the attended motion direction than it would be otherwise. However, in these studies, attention acted to select one of two competing motion stimuli, and thus modified the interaction between processing of these different motion signals. Whether feature-based attention alters the perceived motion direction in the absence of any competing stimuli, has not been tested experimentally.

We conducted a psychophysics experiment to investigate how learned expectations alter perceived motion direction (chapter 3). We found that subjects' learned expectations altered their perception of new stimulus motion directions, inducing an attractive perceptual bias towards frequently presented motion directions. While these perceptual changes occurred as a result of subjects' 'expectations' rather than their 'attentional state' per se, a close relationship is often observed between the perceptual effects produced by both phenomena (Downing, 1988; Pestilli and Carrasco, 2005) (i.e. the perceptual quality of both attended or expected stimuli is increased; see section 3). Our modeling work (chapter 4) lends further support to the hypothesis that attention alters appearance; it predicts that attention should induce attractive perceptual biases towards task-relevant stimuli, in the absence of any changes to the presented stimulus statistics (section 4.2.7).

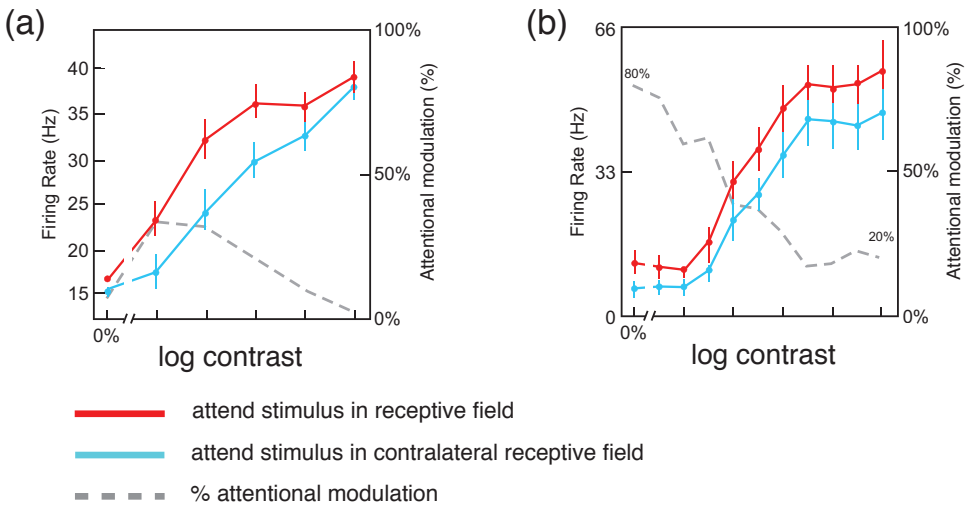


Figure 2.4: Effect of attention on the firing rate of neurons in visual area V4, with varying stimulus contrast (adapted from (Reynolds and Heeger, 2009)). (a) Reynolds et al. found that attention increased neural responses at intermediate, but not at high stimulus contrasts – equivalent to an attention-dependent increase in the effective stimulus contrast (Reynolds et al., 2000). (b) However, Williford & Maunsell found that attention increased neural firing rates at high stimulus contrasts, which could not be explained by an attention-dependent increase in the effective stimulus contrast (Williford and Maunsell, 2006).

2.3.2 Attentional modulation of visual neuron responses

Selective attention modulates neural responses at multiple stages of visual processing (Desimone and Duncan, 1995; Reynolds and Chelazzi, 2004), typically by increasing the firing rate of neurons that are selective to an attended spatial location (Moran and Desimone, 1985; Reynolds et al., 2000) or feature (Spitzer et al., 1988; Treue and Martínez Trujillo, 1999). However, precisely how attentional modulation of neural responses depends on the presented stimulus and behavioural task has been the subject of much debate (Reynolds and Heeger, 2009). Some results, in which attention is found to increase the responses of visual neurons to an intermediate but not a high contrast stimulus (figure 2.4a), are consistent with an attention-dependent shift in neural contrast response functions (neural firing rate plotted against stimulus contrast) (Reynolds et al., 2000; Martinez-Trujillo and Treue, 2002). These results led Reynolds et al. to propose that attention acts as a 'spotlight' that increases the effective contrast of attended stimuli (*contrast gain*) (Reynolds and Chelazzi, 2004; Reynolds et al., 2000). However, other experiments have reported that attention can also increase neural firing rates at high stimulus contrasts, which cannot be explained by an attention-dependent increase in the effective stimulus contrast (figure 2.4b) (Williford and Maunsell, 2006; McAdams and Maunsell, 1999; Motter, 1993). While contradicting the 'contrast gain' principle of Reynolds et al., these experiments are consistent with the proposal that attention increases neural responses multiplicatively, by applying a fixed response gain factor (*response gain*) (McAdams and Maunsell, 1999; Treue and Martínez Trujillo, 1999). Treue et al. proposed that the gain factor depends on the similarity between the neuron's stimulus selectivity, and the location or feature being attended (Treue and Martínez Trujillo, 1999; Martinez-Trujillo and Treue, 2004). They argued that this *feature-similarity gain principle* can account for the observed sharpening of neuronal tuning curves when attention is directed towards a particular feature (Spitzer et al., 1988; Martinez-Trujillo and Treue, 2004). Opposing the idea that attention alters responses via a simple multiplicative gain factor, several studies have reported that the effects of attention are greatly increased when multiple stimuli appear together within a neuron's receptive field ('RF') (Moran and Desimone, 1985; Reynolds et al., 1999). These studies find that, when multiple stimuli are presented within the RF, attention increases or decreases neural responses so that they behave as though only the attended stimulus was present. To explain these results, Duncan & Desimone proposed their *biased competition* theory, in which neurons representing different stimuli compete, and attention biases competition in favour of neurons that encode the attended stimulus (section 2.2) (Desimone and Duncan, 1995).

To account for the diverse effects of attention on the responses of visual neurons, Reynolds & Heeger proposed their *normalization model of attention*, which combines aspects of many of the previous proposals within a single framework (Reynolds and Heeger, 2009). As this model is closely related to our work, we describe it in some detail.

Normalization models of visual neural responses were initially introduced to provide a simple explanation for suppressive phenomena observed in the responses of visual neurons (Simoncelli and Heeger, 1998; Carandini et al., 1997). These models include a 'stimulus drive', representing the excitatory feedforward input to visual neurons, and a 'suppressive drive', representing lateral inhibitory connections. The response of a neuron with RF centred at a location x ('$R(x)$') is computed by dividing the stimulus drive ('$E(x)$') by the suppressive drive ('$S(x)$') plus a constant ('σ') that determines the contrast gain:

$$R(x) = \left| \frac{E(x)}{S(x) + \sigma} \right|_{\mathrm{T}}, \tag{2.1}$$

where $|\cdot|_{\mathrm{T}}$ denotes rectification with respect to a threshold T. The suppressive drive is computed by pooling the stimulus drive from neurons with RFs centred at a range of spatial locations:

$$S(x) = s(x) \star E(x), \tag{2.2}$$

where $\star$ denotes a convolution, and $s(x)$ is the suppressive field, which determines the extent of the spatial pooling.

To provide some intuition about how neural responses are computed from the model, we set the excitatory drive $E(x)$ proportional to the stimulus contrast, c. Thus, the response of a single model neuron can be expressed as a function of contrast:

$$r(c) = \frac{\alpha c}{c + \sigma}, \tag{2.3}$$

where α is a constant of proportionality. From equation 2.3 we can see that when c is small ($c \ll \sigma$), the response grows linearly with contrast, while for large c ($c \gg \sigma$) the response saturates at a fixed value ($r(c) \to \alpha$). Thus, the contrast response function ('CRF') predicted by the model is qualitatively similar to what is observed experimentally, with neural firing rates increasing monotonically at low to intermediate stimulus contrasts, before saturating at high contrasts (figure 2.4).

Reynolds & Heeger extended this normalization model to account for attentional modulation of neural responses. They hypothesized that an 'attention field' ($A(x)$) multiplicatively scales the stimulus drive, which in turn alters the suppressive drive, according to:

$$R(x) = \left| \frac{A(x) E(x)}{S(x) + \sigma} \right|_{\mathrm{T}} \tag{2.4}$$

$$S(x) = s(x) \star [A(x) E(x)]. \tag{2.5}$$

Depending on the relative size of the attended region (determined by the behavioural task), the stimulus drive (determined by the presented stimulus and neural RFs), and the suppressive field (the spatial extent of lateral inhibitory connections), Reynolds & Heeger showed that attention can lead to either a 'contrast gain' or a 'response gain' modulation of neural responses. In the following, we provide a simple explanation of how these effects come about in their model.

First, consider the case where the attention field extends over a much larger region of space than the presented stimulus. In this case, the attention field will be roughly constant in magnitude over the spatial extent of the suppressive field, so that the response of a model neuron is approximated by:

$$r(c) = \frac{\alpha\gamma c}{\gamma c + \sigma} \tag{2.6}$$

$$= \frac{\alpha c}{c + \frac{\sigma}{\gamma}}, \tag{2.7}$$

where $\gamma > 1$ represents the peak of the attention field. Thus, when the attended spatial region is much larger than the stimulus, attention will increase the contrast gain of the model neuron by a factor of γ (so that $\sigma_{eff} \approx \sigma/\gamma$).

Next, consider the case where the attention field extends over a much smaller region of space than the stimulus. In this case, in addition to multiplicatively increasing the stimulus drive at the focus of attention, the spatial extent of the stimulus drive (i.e. the effective stimulus size) will be reduced by attention. Thus, the response of a model neuron with RF centred at the attended location will be approximated by:

$$r(c) = \frac{\alpha\gamma c}{\gamma c + \beta c + \sigma}, \tag{2.8}$$

where $\gamma > 1$ represents the peak of the attention field, and $0 < \beta < 1$ represents the strength of suppression from the region of the stimulus drive outside the focus of attention. When c is small ($c \ll \sigma$), the neural response is approximated by, $r(c) \approx \alpha\gamma c/\sigma$: it is proportional to to the attentional gain, γ. When c is large ($c \gg \sigma$), the neural response saturates at a value given by, $r(c) \approx \alpha\gamma/(\gamma+\beta)$, so that increasing γ will still give rise to an increased response.

To summarize, Reynolds & Heeger's model predicts that when the size of the attentional focus is large relative to the stimulus, a contrast gain effect should be observed, while when the size of the attentional focus is small relative to the stimulus, a response gain effect should be observed. This prediction led Reynolds & Heeger to propose that seemingly contradictory experimental results, where attention gives rise to either a contrast gain (figure 2.4a) (Reynolds et al., 2000) or a response gain (figure 2.4b) (Williford and Maunsell, 2006), could be explained by differences in experimental setup; specifically, due to variations in the relative size of the attended spatial region and the presented stimulus.

In addition to describing attentional modulation of neural CRFs, Reynolds & Heeger's model accounts for the observed effects of attention on neuronal tuning curves (described by the *feature-similarity gain principle* of Treue et al. (Treue and Martínez Trujillo, 1999; Martinez-Trujillo and Treue, 2004)), as well as attentional modulation of neural responses when multiple stimuli are present within the RF (described by Desimone & Duncan's *biased competition* model (Desimone and Duncan, 1995)). *Biased competition* comes about in Reynolds &

Heeger's model as a consequence of divisive normalization, which ensures that neurons representing different stimuli compete. *Response gain* modulation occurs when there is minimal surround suppression, due to the simple multiplicative scaling of the excitatory stimulus drive.

Two alternative models of attention have been proposed that share a similar mathematical form to the normalization model of attention (Lee and Maunsell, 2009; Ghose, 2009). In common with Reynolds & Heeger, Ghose proposed that attention dynamically modulates the inputs to visual neurons, which combine nonlinearly to give rise to the observed neural responses (Lee and Maunsell, 2009). Alternatively, Lee & Maunsell proposed that attention modulates divisive normalization, without affecting the neuronal inputs themselves (Lee and Maunsell, 2009). The merit of these models comes from the fact that they are able to account for a broad range of experimentally observed effects of attention, using a small number of assumptions. However, they make no attempt to explain why attention is required, or how it should be allocated. One might argue that this is not a valid criticism of these models: they were designed to describe the *effects* of attention, not its *cause.* However, the fact that the attentional state is not constrained by the models themselves, makes some of their predictions difficult to test experimentally. For example, as discussed, Reynolds & Heeger's model predicts that the effects of attention on neural CRFs should depend qualitatively on the size of the attentional focus. However, as their model does not describe how the size of the attentional focus is determined by the behavioural task, this prediction will be hard to verify without resorting to vague heuristics about the assumed allocation of attention.

Recently, Chikkerur et al. showed that Reynolds & Heeger's normalization model can be derived from functional principles (Chikkerur et al., 2010), using a normative Bayesian framework. As described previously (section 1.2), Bayesian models of visual processing hypothesize that the brain learns a generative model describing how the hidden state of the world ('$\boldsymbol{s}$') generates the observed sensory input ('$\boldsymbol{I}$'). The assumed goal of visual processing is to invert this generative model, inferring the posterior probability distribution over the hidden states ($p(\boldsymbol{s}|\boldsymbol{I})$). Chikkerur et al. hypothesized that each neuron represents a single binary hidden variable, with firing rate proportional to the posterior probability that the corresponding hidden variable is active: $p(s_i = 1|\boldsymbol{I}) = p(s_i = 1, \boldsymbol{I}) / p(\boldsymbol{I})$. Given certain additional assumptions about the form of the internal model, Chikkerur et al. showed that this expression has a very similar mathematical form to the expression for neural firing rates in Reynolds & Heeger's model. Notably, while divisive normalization was an assumption in Reynolds & Heeger's model, it emerges automatically in the work of Chikkerur et al. as a consequence of the Bayesian formulation of their model (due to the denominator in Bayes' rule, $p(\boldsymbol{I})$). However, in order to simulate the effects of attention, Chikkerur et al. had to impose an ad hoc 'attentional prior' ($p_{att}(\boldsymbol{s})$), analogous to the 'attention field' used by Reynolds & Heeger.

While the model of Chikerrur et al. gives insight into the functional principles that could

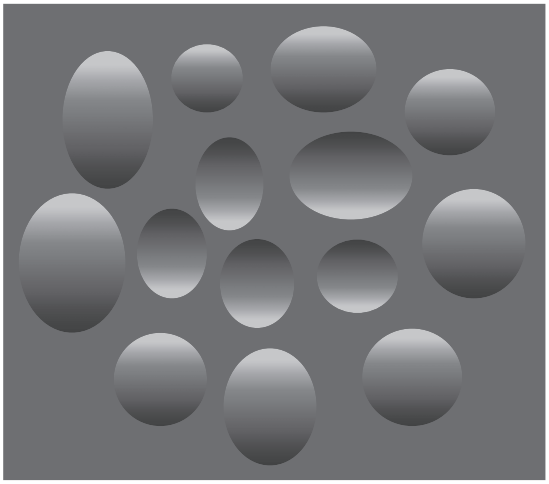


Figure 2.5: Demonstration of the effect of expectations on perceptual appearance. Our strong expectation for light to come from above determines the perceived three-dimensional shape of the ellipses shown above (rotate the page to invert the shapes).

underlie Reynolds & Heeger's normalization model, it suffers from a similar weakness: it does not describe how the 'attentional prior' is determined by behavioural demands and sensory experience. We address this question in chapter 4. In common with Chikerrur et al., our Bayesian model of visual processing gives rise to a similar expression for neural firing rates as Reynolds & Heeger normalization model. However, unlike the work of Chikkerur et al., attention-dependent changes to the perceptual prior are predicted as a direct consequence of optimizing performance in the behavioural task.

Eckstein et al. highlight the difficulty in relating measured changes to neural responses to high-level psychological theories of attention (Eckstein et al., 2009). They use a statistical decision theoretic framework to compare the neural predictions made by two competing theories of attentional selection: a 'limited resources' model, where attention increases sensory sensitivity to a target stimulus; and a 'selective weighting' model, where attention does no alter sensory sensitivity, rather how information is integrated, giving higher weighting to a target stimulus. Interestingly, Eckstein et al. find that the predicted changes to neural activity vary radically depending on their assumptions about the encoded variables, making it very difficult to distinguish between the two theories.

2.4 Similarities & differences between expectations & goal-orientated attention

In chapter 1 I made a distinction between *goal-orientated attention*, which prioritizes perceptual processing of behaviourally relevant stimuli, and *expectations*, which constrain perceptual

processing depending on the statistical likelihood that different stimuli are presented (Summerfield and Egner, 2009). However, this distinction is not always made in the experimental literature, where the perceptual and neurophysiological effects of varying stimulus statistics are often conflated with the effects of varying task context (Posner et al., 1980; Ghose and Maunsell, 2002). A potential reason for this is that attention and expectations are generally thought to be controlled by similar cognitive processes, which allocate increased resources to the perceptual processing of stimuli that are either behaviourally relevant or contextually likely (Corbetta and Shulman, 2002). Recently however, Summerfield & Egner challenged this view, pointing to experimental data indicating that expectations give rise to qualitatively different changes in neural activity from those produced by attention (Summerfield and Egner, 2009). In this section, I briefly review the perceptual and neurophysiological changes that occur as a result of varying the behavioural versus the statistical context of presented stimuli, and discuss the relation between expectations and goal-orientated attention.

2.4.1 Perceptual effects of expectations

It is well established that expectations modulate perceptual performance; for example, by increasing subjects' speed and accuracy at detecting stimuli that are presented at an expected location (Posner et al., 1980; Sekuler and Ball, 1977; Downing, 1988), or by improving the recognition of objects that are expected within the context of a visual scene (Bar, 2004). In addition to modulating perceptual performance, expectations can also alter the subjective appearance of visual stimuli, so that stimuli are perceived as being more similar to what is expected than they actually are (see section 2.3.1 for discussion of the effects of attention on perceptual appearance). These changes in perceptual appearance are strongest when the available sensory inputs are ambiguous; when there are multiple competing explanations for the received sensory input (Haijiang et al., 2006; Sterzer et al., 2008; Adams et al., 2004).

Expectations can be manipulated quickly, through instructions (Sterzer et al., 2008), sensory cues (Posner et al., 1980), or exposure to novel stimulus statistics (Sotiropoulos et al., 2011). However, in addition to rapidly acquired expectations that reflect the current stimulus context, people also exhibit global expectations that reflect the statistical structure of natural images (Sekuler and Ball, 1977; Posner et al., 1980; Downing, 1988). For example, our expectation for light to come from above determines how we perceive the shaded circles shown in figure 2.5 (Sun and Perona, 1998). This type of expectation is presumably acquired over long periods of time, during evolution and development, and thus might be assumed to be resistant to change (Hyvärinen, 2010; Geisler, 2003). Interestingly however, recent work has shown that long-term expectations can be altered or reversed as a result of acquired sensory experience, indicating that visual expectations are continuously updated in the light of new information about the environment (Adams et al., 2004; Sotiropoulos et al., 2011).

2.4.2 Neurophysiological effects of expectations

Expectations and attention modulate perceptual performance in a qualitatively similar way, increasing detection performance and accuracy in discriminating stimuli that are either attended or expected. Therefore, one might anticipate that they would give rise to similar neurophysiological changes. However, a recent review by Summerfield & Egner suggests that this is not the case (Summerfield and Egner, 2009). While attention is consistently observed to increase visual responses to an attended stimulus (Reynolds and Chelazzi, 2004), Summerfield & Egner cite fMRI and EEG data reporting a reduction in visual responses are reduced towards stimuli that are statistically likely (Yoshiura et al., 1999; Marois et al., 2000).

Summerfield & Egner argue that the observed effects of expectations on neural responses are consistent with *predictive coding*: the hypothesis that sensory neurons a 'prediction error', relating to the difference between the sensory input that is expected and the sensory input that is received (Rao and Ballard, 1999; Schultz and Dickinson, 2000). If this hypothesis is true, then an expected stimulus should produce a smaller prediction error, giving rise to a reduction in neural firing rates. In other words, as the predictive coding framework assumes that the visual system is set up to process *unexpected input* (i.e. surprises), goal-directed attention and expectations should produce very different changes to visual processing: the visual system should be desensitized to expected (i.e. unsurprising) sensory inputs, but sensitized to 'interesting' (i.e. task-relevant or surprising) sensory inputs.

Unfortunately, there has been little experimental investigation into how expectations alter the activity of single neurons (Summerfield and Egner, 2009). However, there is some evidence to suggest that, contrary to the imaging data reviewed by Summerfield & Egner, the responses of visual neurons are *increased* when the stimuli that they encode are expected (Ghose and Maunsell, 2002; Ghose and Bearl, 2010; Jaramillo and Zador, 2011). For example, Ghose et al. conducted experiments investigating how the responses of neurons in V4 and MT depend on the precise temporal structure of a behavioural task (Ghose and Maunsell, 2002; Ghose and Bearl, 2010). They used an experimental protocol where monkeys had to detect a stimulus change (a change in luminance) whose probability of occurrence varied in time. Attention-dependent increases in the responses of neurons encoding task-relevant stimuli were found to depend on the probability of stimulus change at each moment in time; largest increases in response were observed when the probability of stimulus change was high.

Further experimental evidence suggesting that expectations increase neural firing rates comes from experiments reporting that the responses of visual neurons are increased when the stimuli that they encode are 'expected' from their contextual surroundings (Komatsu, 2006). For example, the responses of V1 neurons to a low contrast orientated stimulus presented within their RF are increased by the addition of collinear flankers presented in the surrounding region of space (Kapadia et al., 2000; Polat et al., 1998) (although the opposite effect is observed at

high-contrast; see (Seriès et al., 2003) for discussion of experimental controversy).

How can these conflicting experimental data be reconciled? One possibility is that the qualitative effect of expectations on neural activity depends on the behavioural relevance of expected stimuli; sensory signals produced by a frequently observed but behaviourally irrelevant stimulus could be filtered out during sensory processing, while sensory signals that come from an expected stimulus that is behaviourally relevant are enhanced. For example, an experimental paradigm where expectations are consistently found to decrease visual responses is the 'odd-ball' task, in which subjects are required to detect a rare (i.e. unexpected) target stimulus presented amongst frequent (i.e. expected) distractors (Yoshiura et al., 1999). On the other hand, Ghose et al. found that the responses of visual neurons are increased when a behaviourally relevant change in the encoded stimulus is expected (Ghose and Maunsell, 2002).

If the effects of expectations depend on behavioural demands, then it may be hard to distinguish the effects of goal-orientated attention from the effects of expectations. For example, in our psychophysics task, subjects' expectations were manipulated by presenting stimuli moving in some directions more frequently than other directions. However, as subjects were required to perform a behavioural task, reporting the direction that stimuli were moving in, their perception of presented stimuli could have been altered by the behavioural relevance, as well as the statistical likelihood of presented stimuli. Nevertheless, a simple Bayesian model, that assumed that subjects' perception of presented stimuli was altered by learned information about the stimulus statistics alone was able to provide a good description of their behaviour in this task.

In chapter 4 we propose a Bayesian model of visual processing that can account for changes in visual neural responses that occur either due to changes in either the behavioural task or the presented stimulus statistics. In our simulations we consider an experimental protocol in which all stimuli are equally likely to be presented but only certain stimuli are relevant to the task. In this case, and given certain assumptions about the observer's internal model and neural code, our model predicts that the firing rate of neurons that are tuned to behaviourally relevant stimuli should be increased. While the simulated protocol was chosen deliberately to investigate the neural effects of varying behavioural context in the absence of varying stimulus statistics, in the future it would be interesting to investigate the case where both the task demands and stimulus statistics are manipulated simultaneously, as is often the case experimentally.

2.4.3 Bayesian formulation of expectations and attention

We might try to use the available neurophysiological and perceptual data to ask: "what is the relation between expectations and goal-orientated attention?" However, this question is poorly defined: there is no reason to assume that a unitary cognitive process associated with either 'expectations' or 'attention' exists. Furthermore, the perceptual and neurophysiological effects of

these phenomena are complex and depend nontrivially on the precise setup of the behavioural task and the presented stimuli. Alternatively, rather than getting caught in the (possibly false) dichotomy between 'expectations' and 'attention', we can choose to abandon these terms altogether, asking the more concrete question: "how do changes in behavioural demands and stimulus statistics alter visual processing?" In this view, we would not consider expectations and attention as the *cause* of changes in visual processing, but see them as descriptive terms, referring to the perceptual and neurophysiological *consequences* of changing stimulus statistics and behavioural demands (Anderson, 2011).

Note this view is not incompatible with an internal mechanism that causes attention-dependent changes in perception and neural responses. However, as long as we do not know what this internal mechanism is, it makes sense to consider the 'higher cause' of such internal changes which are under direct experimental control: namely, the presented stimuli and behavioural task.Bayesian models provide a natural framework for considering how the statistics of sensory stimuli modulate perceptual processing. In these models, subjects' expectations are represented by a prior probability distribution, which denotes the probability that they associate with different states of the world. Bayesian theory describes how prior expectations should be combined with sensory signals to perform perceptual inference, as well as how the prior should be updated when new sensory information is received (Jaynes, 1986; MacKay, 2003). Thus, Bayesian models make precise predictions about the perceptual biases and changes in perceptual performance that should be induced by a given perceptual prior.

These predictions have been tested in numerous psychophysical experiments, which indicate that in simple tasks, people combine their prior expectations with available sensory evidence in a probabilistically optimal manner (Stocker and Simoncelli, 2006b; Weiss et al., 2002; Girshick et al., 2011; Knill, 2007; Körding and Wolpert, 2004). Some researchers have investigated the effect of 'structural priors', that reflect the statistics of natural sensory signals. For example, Simoncelli and colleagues have shown that people exhibit systematic biases in their estimates of visual features such as orientation and motion direction, which can be explained by assuming that they use a perceptual prior that is well matched to the statistics of natural sensory signals (Weiss et al., 2002; Stocker and Simoncelli, 2006b; Girshick et al., 2011). Other researchers have investigated the effect of 'contextual priors', that reflect the stimulus statistics in a particular experimental context. These researchers have shown that people are able to learn prior expectations about novel statistics introduced during a psychophysical task, and that they combine these expectations with available sensory information in a manner consistent with Bayesian inference (Adams et al., 2004; Körding and Wolpert, 2004; Knill, 2007; Sotiropoulos et al., 2011).

However, previous experiments looking at rapidly learned 'contextual priors', provided subjects with additional (haptic) feedback during learning, which could be used to disam-

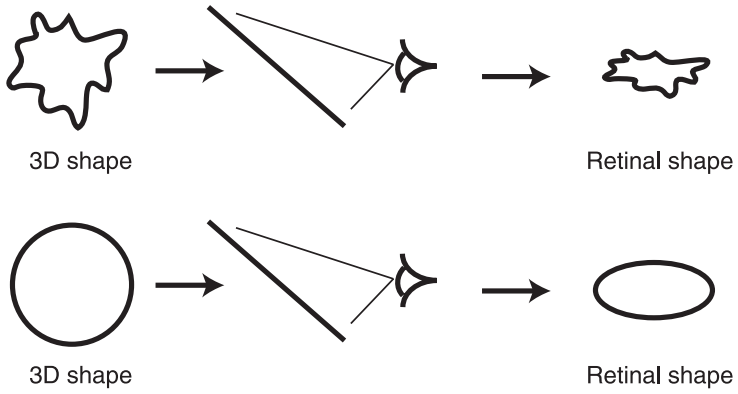


Figure 2.6: When viewed at a slant, planar figures (left) project to compressed figures in the retinal image (right). A statistical tendency for figures to be isotropic (distributed evenly in all directions) means that the projected retinal image is informative about 3D surface orientation. Thus, people are more likely to interpret an elliptical retinal image as corresponding to an oblique circle than a vertical ellipse.

biguate the presented visual stimuli (Adams et al., 2004; Körding and Wolpert, 2004). As, in most real-world situations people do not receive haptic information about their received visual input, an important question is whether sensory priors can be acquired quickly from visual input alone. To address this question, Knill conducted a psychophysics experiment investigating how subjects' learned expectations about stimulus shape alter how they interpret pictorial cues to depth (Knill, 2007). In this experiment, subjects were asked to judge the planar orientation of randomly shaped ellipses (figure 2.6). Under normal conditions, subjects exhibited a prior expectation for regularly shaped objects, causing elliptical stimuli to be perceived as circles presented at an oblique angle. Prolonged exposure to a stimulus distribution that included a large number of randomly shaped ellipses reduced subjects' prior expectation for circular stimuli. Consequently, after training, they gave progressively less weight to stimulus shape, and more weight to stereoscopic cues, in their estimates of stimulus slant.

In Knill's experiment, subjects' learned expectations influenced how they combined different sources of sensory information (pictorial versus stereoscopic cues): subjects learned that the stimulus shape was an unreliable cue for judging the stimulus slant, causing them to rely more strongly on stereoscopic cues. However, as well as altering how different sources of sensory information are combined, subjects' learned expectations should alter their perception of simple stimulus features, so that they appear more similar to expected stimulus features than they actually are. To test this prediction, we examined whether expectations of simple stimulus features can be developed implicitly through a fast statistical learning procedure, and if so, how these learned expectations bias subjects' perception of newly presented stimuli (chapter 3). We found that subjects rapidly learned to expect frequently presented motion directions, and that

these learned expectations resulted in estimation biases towards these motion directions, as well as hallucinations when no stimulus was presented. Subjects' behaviour in the task was well explained by a model that assumed that they followed a Bayesian strategy, combining their learned expectations (the prior) with received sensory data (the likelihood) according to Bayes' law.

While Bayesian theory makes unambiguous predictions about how prior expectations should alter visual perception, the predicted changes to neural responses are less clear. One reason is that it is not known how (or whether) probability distributions are encoded by neurons (section 5.3). Nonetheless, given certain assumptions about the encoding scheme, some researchers have investigated how prior expectations should alter sensory neural responses (Ganguli and Simoncelli, 2010; Shi and Griffiths, 2009; Simoncelli, 2009). Indeed, many of the Bayesian models of attention discussed earlier attribute the effects of attention on neural responses as due to changes in the perceptual prior (Dayan and Zemel, 1999; Yu and Dayan, 2005a; Rao, 2005; Chikkerur et al., 2010) (see section 2.1). In these models, the main effect of increasing the prior probability associated with a particular stimulus feature or location is to increase the firing rate of model neurons that encode this feature or location. However, in addition to depending on the assumed encoding scheme, the effect of varying stimulus statistics on neural responses will also depend on the structure of the internal model in the brain (section 5.1). For example, Schwartz and others have shown that, if the aim of visual processing is to extract 'independent' components of sensory signals, presenting a particular stimulus frequently could give rise to a reduction in the responses of neurons that are tuned towards this stimulus (Wainwright et al., 2001; Schwartz and Simoncelli, 2001; Schwartz et al., 2007, 2009). Thus, understanding how expectations alter neural responses requires asking deep questions about how sensory information is represented in the brain, and how this information is encoded neurally.

The behavioural relevance of sensory signals depends on how useful they are in making decisions, or performing actions. Thus, to understand how behavioural demands should alter perceptual processing we can consider a Bayesian decision theory framework, which describes how uncertain sensory information should be used to make perceptual decisions (Yuille and Bulthoff, 1996; Körding and Wolpert, 2006). Behavioural demands are incorporated in this framework through the use of a utility function, which denotes the expected utility associated with each decision, given the hidden state of the world (equation 1.2). However, while Bayesian decision theory predicts how behavioural demands should influence *decisions*, it does not necessarily follow that behavioural demands should also influence perceptual processing itself. That is, if there are no constraints on perceptual processing and the agent is able to learn the 'true' model of their environment, the inferred posterior distribution should not be affected by behavioural demands: if perceptual processing is unlimited, *all* sensory signals should be processed, not just those that are behaviourally relevant.

Understanding how behavioural demands influence perception requires considering the limited computational resources that are available to process sensory signals (discussed in section 2.1). In contrast, varying the stimulus statistics should alter perceptual processing even in the absence of any limited computational resources, to improve inferences about intrinsically ambiguous sensory information. In general, the visual system is faced with both *external* limitations on the available sensory information, and *internal* limitations on the computational resources available to process this information. Together, expectations and attention describe the strategies that the brain uses to deal with these limitations; where prior knowledge about the statistics and behavioural relevance of sensory signals is used to optimize perceptual processing and improve behavioural performance.

Chapter 3

Effect of learned expectations on visual motion perception

In this chapter, we describe a psychophysics experiment that was conducted to investigate whether expectations of simple stimulus features can be developed implicitly through fast statistical learning, and if so, how these expectations are combined with visual signals to modulate perception. We examined this question in the context of motion perception, in a design where some motion directions were more likely to appear than others. Our hypothesis was that participants would automatically learn which directions were most likely to be presented and that these learned expectations would bias their perception of motion direction. A secondary hypothesis was that participants would solve the task using a Bayesian strategy, combining a learned prior of the stimulus statistics (the expectation) with their sensory evidence (the actual stimulus) using Bayes' rule. This work was published in Journal of Vision (Chalk et al., 2010), and the main findings were replicated by Gekas et al. (Gekas et al., 2011).

3.1 Methods

3.1.1 Observers and stimuli

Twenty naive observers with normal or corrected-to-normal vision participated in this experiment. All participants in the study gave informed written consent, received compensation for their participation and were recruited from the Riverside, CA area. The University of California, Riverside Institutional Review Board approved the methods used in the study, which was conducted in accordance with the Declaration of Helsinki.

Visual stimuli were generated using the Matlab programming language and displayed using Psychophysics Toolbox (Brainard, 1997; Pelli, 1997) on Viewsonic P95f monitor running at 1024X768 at 100Hz. The display luminance of the CRT monitor was made linear by means of an 8-bit lookup table. Participants viewed the display in a darkened room at a viewing distance

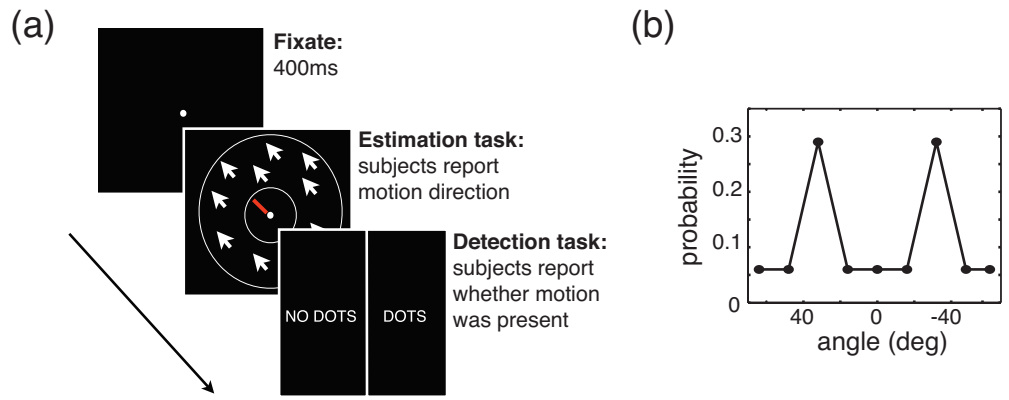


Figure 3.1: (a) Sequence of events in a single trial. Each trial began with a fixation point, followed by the appearance of a motion stimulus. A central bar projecting from the fixation point was presented simultaneously with the motion stimulus, and allowed participants to estimate the direction of motion. After either participants had made an estimation, or a period of 3000ms had elapsed, the stimulus disappeared and was replaced by a vertical line, with text to either side. Participants moved a cursor to either side of the line to indicate whether they had perceived the motion stimulus. (b) Probability distribution of presented motion directions. Two directions, 64° apart from each other, were presented in a larger number of trials than other directions. Motion direction is plotted relative to a reference direction at 0°, which was different for each subject.

of 100 cm with their motion constrained by a chin rest. Motion stimuli consisted of a field of dots (density: 2 dots/deg^2 at 100Hz refresh rate) moving coherently at a speed of 9°/sec within a circular annulus, with minimum and maximum diameter of 2.2° and 7° respectively. The background luminance of the display was set to 5.2 cd/m^2.

3.1.2 Procedure

At the beginning of each trial a central fixation point (0.5° diameter, luminance 12.2cd/m^2) was presented for 400 ms. With the fixation point still onscreen, the motion stimulus was then presented, along with a red bar which projected out (initial angle of bar randomized for each trial) from the fixation point (figure 3.1a). The bar was located entirely within the centre of the annulus containing the moving dots (length 1.1°, width 0.03°, luminance 3.4cd/m^2). Participants indicated the direction of motion by orienting the red bar with a mouse, clicking the mouse button when they had made their estimate (estimation task). The display cleared when either the participant had clicked on the mouse, or a period of 3000ms had elapsed. On trials where no motion stimulus was presented, the red bar still appeared and participants were required to estimate the perceived direction of motion as normal. Participants were instructed to fixate on the central point throughout this period. Participants' reaction time in the estimation task

determined how long the stimulus was presented for. On average this was equal to 1978±85ms (standard error on the mean; see figure 3.10c for a plot of reaction time versus presented motion direction).

Note that, while participants' were requested to maintain fixation throughout the period while the motion stimuli was displayed, we did not check whether they maintained fixation on the centre of the screen. Thus it is possible that participants' eye-movements could have influenced their behaviour in the task (see section 3.4.4 for discussion).

After the estimation task had finished, there was a 200ms delay before a vertical white line was presented at the centre of the screen, with text to either side (reading 'NO DOTS' and 'DOTS' respectively). Participants moved a cursor to the right or left of this line to indicate whether they had or had not seen a motion stimulus, and clicked the mouse button to indicate their choice (detection task). The cursor flashed green or red for a correct or incorrect detection response, respectively. The screen was then cleared and there was a 400 ms blank period before the beginning of the next trial.

Every 20 trials, participants were presented block feedback on the estimation task, with text display on screen informing participants their average estimation for the previous 20 trials (e.g. "In the last 20 trials, your average estimation error was: 20°"). Block feedback, rather than trial-by-trial feedback was given, because we wanted to encourage participants to do their best at the estimation task, without interfering with their estimation behaviour on each trial.

After completing both experimental sessions, participants were handed a questionnaire, where they were asked to comment on the stimuli presented during the experiment, and in particular, whether they were aware of stimuli moving in some directions more than others (figure 3.2).

3.1.3 Design

Participants took part in two experimental sessions lasting around one hour each, taken over successive days. Each session was divided into 5 blocks of 170 trials where all stimulus configurations were presented, making 1700 trials in total (850 trials per session).

Participants were presented with stimuli at 4 different randomly interleaved contrast levels. The highest contrast level was at 1.7cd/m^2 above the 5.2cd/m^2 background. For each session there were 250 trials at zero contrast and 100 trials at high contrast. Contrasts of other stimuli were determined using a staircase procedure on subjects' detection performance (García-Pérez, 1998). For each session there were 135 trials using a 2/1 staircase and 365 trials using a 4/1 staircase (an $n/1$ staircase implies that the stimulus contrast decreases by a fixed step-size following n correct detection responses, and increases by the same amount following a single incorrect response).

For the two staircased contrast levels, on a given trial the direction of motion could be

Questionnaire

1. Did you find the first or second easier, or were they both the same? (please circle as appropriate)

   1st session easier 2nd session easier about the same

2. Did you notice anything unusual about the number of motion stimuli that were moving in each direction? For example, were some directions shown more than others? If yes, please describe in more detail what you saw.

3. If you filled in the last question, then was this the same in both sessions? (circle as appropriate)

   Yes No Don't know

4. How many directions of motion do you think there were? (circle as appropriate)

   1, 2, 3, between 4 and 10, more than 10, don't know

5. Did you ever think you saw moving dots, and then it turned out there were none there?

   Yes No Don't know

6. Which of the following descriptions best describes the distribution of motion directions that you saw? (tick statement that you most agree with)

   (a) There were equal number of stimuli moving in all directions.

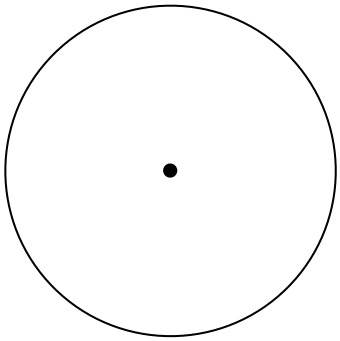
   (b) Most of the stimuli were centred around one central direction of motion.

   (c) Most of the stimuli were centred around 2 different directions of motion.

   (d) There were only two possible directions of motion.

   (e) Don't know.

7. If in the last question you selected (b), (c) or (d), can you draw a line (or lines) from the centre of this diagram out to the edge, indicating the direction(s) that were most frequently presented?



1

Figure 3.2: Experimental questionnaire. The form was completed by participants after completing their second experimental session.

0° ±16°, ±32°, ±48° or ±64°, with respect to a central reference angle. To reduce potential biases in the population averaged results due to reference repulsion from cardinal motion directions (Rauber and Treue, 1998), this central motion direction was randomized across participants. We manipulated participants' expectations about which motion directions were most likely to occur by presenting stimuli moving at ±32° more frequently than the others (figure 3.1b). Therefore, at the 4/1 staircased contrast level, there were 130 trials per session with motion at -32° and +32°, and 15 trials per session for each of the other directions of motion. At the 2/1 staircased contrast level there were an equal number of stimuli moving in each of the predetermined directions: 15 trials per session for each motion direction. At the highest contrast level there were 25 trials per session with motion at -32° and +32° and 50 trials per session at completely random directions (among all possible directions, not just the predetermined directions used in the rest of the experiment).

3.1.4 Data analysis

In our analysis of the estimation task, we looked only at trials where participants both reported seeing a stimulus and clicked on the mouse during stimulus presentation to indicate their estimate of motion direction. The first 100 trials from each session (~25 trials from each contrast staircase) were excluded from analysis, to allow the staircases to converge on stable contrast levels. Data was analyzed for the 12 (of 20) participants who could adequately perform both tasks according to our predetermined performance criteria of detection performance greater than 80% and mean absolute estimation error less than 30° with the highest contrast stimuli in both experimental sessions (see section 3.2.1 for analysis of excluded participants' estimation performance). Importantly, our analysis of participants' performance in the estimation task looked only at their responses to staircased contrast levels, and not their responses to the highest contrast stimuli, which we used to determine which participants should be included.

In the estimation task, the variance of participants' motion direction estimates tended to be quite large and varied greatly across different participants and motion directions. We postulated that this was due to the fact that in some trials participants made completely random estimates. Thus, we fitted participants' estimation responses to the distribution:

$$p(\theta_{est}|\mu,\kappa,a) = (1-a)\,\mathcal{V}(\theta_{est};\mu,\kappa) + \frac{a}{2\pi}, \tag{3.1}$$

where 'a' denotes the proportion of trials where the participant make random estimates, and '$\mathcal{V}(\theta_{est};\mu,\kappa)$' is a von Mises (circular normal) distribution with mean 'μ' and width determined by '1/κ', given by:

$$\mathcal{V}(\theta_{est};\mu,\kappa) = \frac{1}{2\pi I_0(\kappa)} \exp(\kappa\cos(\theta_{est}-\mu)), \tag{3.2}$$

where $I_0(\kappa)$ is the modified Bessel function of order 0. Parameters (a, κ and μ) were fitted for

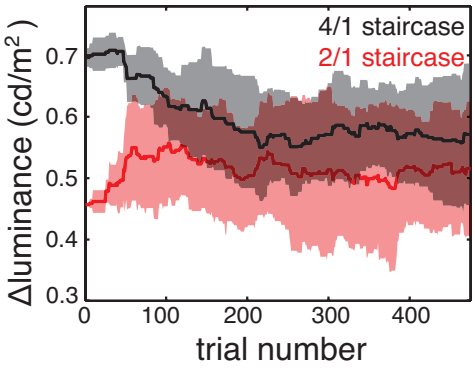


Figure 3.3: Population averaged stimulus contrast, relative to background contrast, for the 4/1 (blue) and 2/1 (red) staircased contrast levels, plotted against trial number (from the 1st experimental session only).

each presented motion direction separately, to maximize their log-likelihood:

$$\{a,\kappa,\mu\} = \arg\max_{\{a,\kappa,\mu\}} \sum_i \log p\left(\theta^i_{est}|\mu,\kappa,a\right), \tag{3.3}$$

where the summation is taken over all trials with a particular presented motion direction. Participants' estimation mean and standard deviation were taken as the circular mean and standard deviation of the von Mises distribution (μ and σ respectively) The mean estimation biases obtained using this method were qualitatively similar to those obtained by simply averaging across trials. However, the variances obtained from the parametric fits were significantly smaller and more consistent across participants, than when the estimation variance was measured directly from the data. Therefore, in all of the following analysis we used this parametric method to quantify estimation biases and variances.

There was no significant interaction between experimental session and motion direction on the estimation bias or standard deviation (p = 0.11 and p = 0.41 respectively, 4-way within-subjects ANOVA). Therefore, we collapsed data across the two experimental sessions.

There was a considerable degree of overlap between the luminance levels achieved using both staircases (figure 3.3). After discounting the first 100 trials from each session, the population averaged standard deviation in the luminance of the 2/1 and the 4/1 staircased levels over the course of one experimental session was 0.051±0.001cd/m^2 and 0.054±0.001cd/m^2 respectively; similar to the average luminance difference between the two levels (0.052±0.004cd/m^2). Further, there was no significant difference between the luminance levels achieved for both staircases (p = 0.23, 3-way within-subjects ANOVA). This was reflected in the estimation data: there was no significant difference between participants' estimation standard deviations for both staircased contrast levels (p = 0.12, 4-way within-subjects ANOVA). Therefore, we collapsed data across these contrast levels for all of the analysis described in the main text (al-

though see section 3.2.3 for our analysis of how the stimulus contrast influenced participants' estimation behaviour).

To analyze the distribution of estimations when no stimulus was present, we constructed histograms of participants' responses, binned into 16° windows. We converted these response histograms into probability distributions, by normalizing them over all motion directions for each participant individually. There was no significant interaction between experimental session and motion direction on the response histograms ($p = 0.87$, 4-way within-subjects ANOVA). There was also no significant 3-way interaction between motion direction, experimental session and detection-response ($p = 0.81$, 4-way within-subjects ANOVA). Therefore we collapsed data across experimental sessions for analysis of the participants' responses when no stimulus was present.

We were interested in how the uneven distribution of presented motion directions influenced participants' perception of the motion stimuli. By design, the probability distribution of presented motion stimuli was symmetrical around a central motion angle (figure 3.1b). Therefore, we reasoned that any asymmetry in participants' estimation and detection behaviour for stimuli moving to either side of the central motion direction was likely due to factors other than the distribution of presented stimuli that was used, such as 'reference biases' towards or away from caudal motion directions (Rauber and Treue, 1998; Girshick et al., 2011). Rather than investigating these systematic biases directly, we attempted to average out their effect, by averaging data from participants presented with different stimulus distributions (i.e. with a different central motion direction), as well as averaging data from when stimuli were moving to either side of the central motion direction. For the estimation task, this also required reversing the sign of the estimation biases for stimuli moving anti-clockwise from the central motion direction before averaging (see appendix A for 'unfolded' versions of figures 3.5a, 3.7a & 3.8).

3.2 Results

3.2.1 Performance of subjects in detection and estimation task

In order to ensure that participants performed adequately in the psychophysical task we used a predetermined performance criteria for inclusion into the study. First, participants were required to detect the motion stimuli on more than 80% of trials with the high contrast motion stimuli while also making active estimates of the motion directions by clicking the mouse. Second, their average estimation performance on the high contrast stimuli had to be within 30° of the correct angle.

We discounted 3/20 participants who did not meet our first criterion in either experimental session. The included participants managed to both detect stimuli and click on the mouse during stimulus presentation to make an estimation of motion direction, on almost every trial

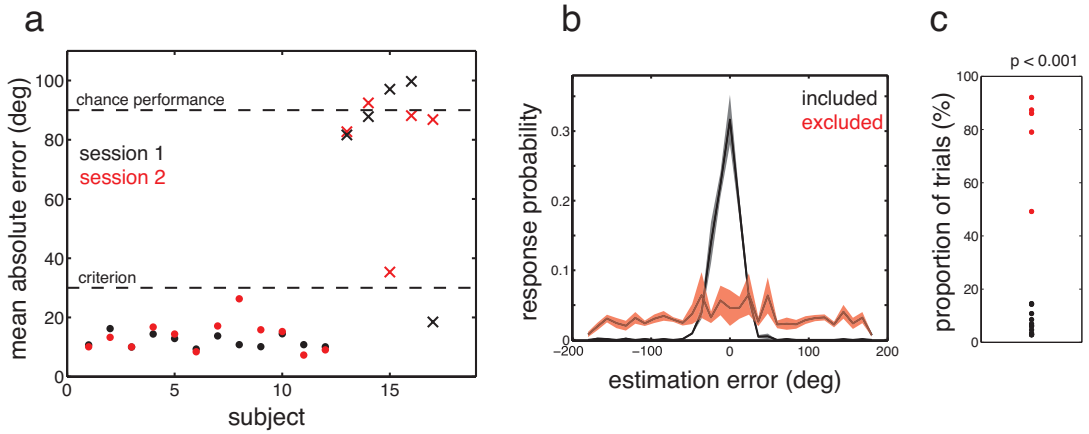


Figure 3.4: Performance of different participants in estimation task, with the high contrast stimuli. (a) The mean absolute estimation error is plotted separately for each experimental session (session 1 and 2 are plotted in blue and red respectively), and for each participant. Participants whose rms estimation error was less than 30° in both sessions were included in our analysis, and are denoted by filled dots while participants who did not meet this criterion were discounted from our analysis are denoted by crosses. The mean absolute error that corresponds to chance performance in the task (90°) and our criterion rms error (30°) are denoted by horizontal dashed lines. (b) Response probability histogram of estimation error with the high contrast stimuli, for included (red) and excluded participants (red). (c) Fraction of trials where participants moved the bar less than 1° from its initial position during the estimation task. Included and excluded participants are shown in red and black respectively.

with the high contrast stimuli (97±0.3% of trials).

The 17/20 participants who passed the criterion for the detection task could be separated according to their estimation performance into two distinct groups (figure 3.4a): 12/20 participants who passed our criterion and performed well in the estimation task (population averaged absolute error of 12.8±0.9°) and 5/20 participants who failed our criterion for the estimation task, performing at near chance levels (with an average rms error of 77.0±4.9°, compared to an average absolute error of 90° that would be expected if they made completely random estimations). Figure 3.4b illustrates the estimation error response probability histograms for included participants (blue) and excluded participants (red) in response to the high contrast stimuli. It is clear from this plot that the excluded participants performed extremely badly at the estimation task, with a distribution of estimation errors that was almost uniform ($p = 0.19$, 2-way within-subjects ANOVA), even with the highly visible high contrast stimuli.

If excluded participants really were not attempting the estimation task at all, then we thought it likely that they would click on the bar immediately during the estimation task, without moving it from its initial (random) orientation. This is indeed what we found: on average the excluded participants did not move the bar more than 1° from its initial position on 79±5% of trials with the high contrast stimuli; significantly more than 7±1% of trials for included participants ($p<0.001$ rank-sum test; figure 3.4c). Excluded participants also performed the estimation task more quickly than included participants, further supporting the argument that they were not really trying to do well in this task (average reaction time of 1.44±0.07s as opposed to 0.89±0.12s for the included versus the excluded participants; $p = 0.027$, rank-sum test).

Our results suggest that rather than just performing worse in the estimation task due to finding it difficult, excluded participants did not try to perform the estimation task at all: they left the estimation bar in its initial position and performed at near chance levels, even with the highly visible high contrast motion stimuli.

3.2.2 Estimates of motion direction when no stimulus present

We investigated whether participants learned to expect the most frequently presented motion directions. To assess this, we analysed participants' estimation responses on trials where no stimulus was presented, but where they reported seeing a stimulus in the detection task, as well as clicking on the mouse to estimate its direction. On average this occurred on 46±3 trials for each participant (10.8±2% of the total number of trials where no stimulus was presented). For this subset of trials, participants' estimation response probability varied significantly with motion direction, with a clear peak close to the most frequently presented motion directions (±32°; $p < 0.001$, 3-way within-subjects ANOVA; figure 3.5a, grey). We quantified the probability ratio that participants made estimates that were close to the most fre-

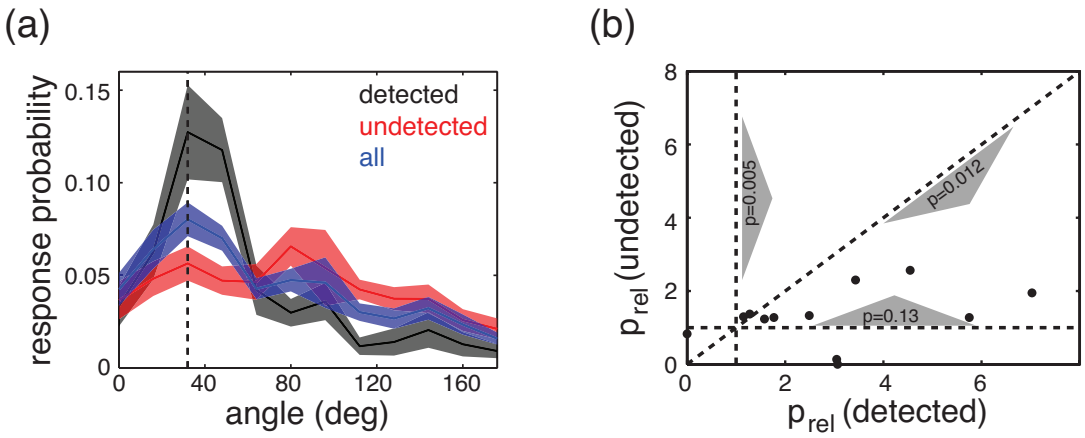


Figure 3.5: Estimation responses in the absence of a stimulus. (a) Probability distribution of participants' estimates of motion direction when no stimulus was present. Response distributions are plotted for all trials (blue), as well as the subset of trials where participants reported detecting a stimulus (grey) and trials where they didn't (red). Data points from either side of the central motion direction have been averaged together in this plot, so that the furthest left data point corresponds to the central motion direction, and the vertical dashed line corresponds to the most frequently presented motion directions (±32°). Results are averaged over all participants and error bars represent within-subject standard error. (b) Probability ratio (p_{rel}) that individual participants estimated within 8° from the most frequently presented motion directions (±32°) relative to other 16° bins, plotted for trials where the stimulus was undetected versus trials where the stimulus was detected. p_{rel} was significantly greater than 1 for trials where participants reported detecting stimuli (p = 0.005, signed rank test), but was only marginally so when subjects failed to detect the stimulus (p=0.13). Participants were also significantly more likely to estimate in the direction of the frequently presented motion directions on trials where they reported detecting stimuli, versus trials where they did not (p = 0.012).

quently presented motion directions, relative to other directions, by multiplying the probability that they estimated within 8° of these motion directions by the total number of 16° bins: $p_{rel} \equiv p\,(\theta_{est} = \pm 32\,(\pm 8)^{\circ}\,|\text{detected}) \times N_{bins}$. This probability ratio would be equal to 1 if participants were equally likely to estimate within 8° of ±32° as they were to estimate within other 16° bins. We found that the median value of p_{rel} was significantly greater than 1, indicating that participants were biased to report motion in the most frequently presented directions when no stimulus was presented (median(p_{rel}) = 2.7; p = 0.005, signed rank test, comparing p_{rel} to 1; figure 3.5b).

As on a large proportion of trials the presented motion stimuli were moving in one of two directions, it is possible that participants could have habituated to automatically move the estimation bar towards one of these two directions, irrespective of their response in the detection task (note that the initial bar position was randomized on each trial and thus biases can't arise from just leaving the mouse in its initial location). In this case we would also expect their 'no-stimulus' estimation distributions to be biased towards the two most frequently presented directions for trials where they did not detect a stimulus. However, on trials where participants did not report seeing a stimulus in the detection task (but where they did click the mouse while the stimulus was present to estimate its motion direction; on average this occurred on 134±9 trials for each participant; 32±7% of the total number of trials where no stimulus was presented), there was no significant variation in the estimation response probability with motion direction (p = 0.12, 3-way within-subjects ANOVA; figure 3.5a, red). Further, for these trials, participants were not significantly more likely to estimate close to the most frequently presented motion directions than other motion directions (median(p_{rel}) = 1.28; p = 0.13, signed rank test, comparing p_{rel} to 1; figure 3.5b). Indeed they were significantly more likely to report motion in the most frequently presented motion directions when they also reported detecting a stimulus, compared to when they did not (p = 0.012, signed rank test, comparing the values of p_{rel} obtained for trials where participants either did or did not report seeing a stimulus in the detection task; figure 3.21b).

An alternative possibility that could produce similar results, is that participants' expectations influenced their behaviour in the detection task but not in the estimation task. Thus, in the absence of a presented stimulus, they would be more likely to report detecting a stimulus when they mistakenly perceived motion in one of the two most frequently presented motion directions, although their estimation responses would be unaltered by their expectations. In this case, participants' estimation responses would be distributed uniformly when we looked at data from all trials where no stimulus was presented (regardless of their response in the detection task). This was not what we found: when we looked at data from all zero-stimulus trials, participants estimation response probability varied significantly with motion direction (p<0.001, 3-way within-subjects ANOVA; figure 3.5a, blue) and they were biased to report

motion in the two most frequently presented directions (median(p_{rel}) = 1.71; $p < 0.001$, signed rank test comparing p_{rel} to 1). However, the size of this bias was reduced, compared to the case when we looked only at trials where participants detected stimuli ($p = 0.027$, signed rank test comparing the values of p_{rel} obtained for all trials with trials where participants reported seeing a stimulus in the detection task).

A final possibility is that, when participants were uncertain about the stimulus motion direction, they made estimations that were influenced by the stimulus presented immediately beforehand. In this case, we would expect the observed biases in participants' no-stimulus estimation distributions to disappear when we excluded trials that were immediately preceded by stimuli moving in the most frequently presented directions (±32°). However, when we excluded these trials from our analysis, participants' zero-stimulus estimations (for trials where they reported detecting a stimulus) were still strongly biased towards the two most frequently presented directions (median(p_{rel}) = 2.11; $p = 0.026$, signed rank test, comparing p_{rel} to 1). It should be noted, nonetheless, that there is a continuum between expecting the next stimulus to be the same as the previous stimulus, and having an expectation based on experience of further back in the past. While our results rule out the most trivial possibility – that participants' expectations are influenced only by the immediately presented stimulus – further work would be required to understand exactly how people integrate information about previously presented stimuli to inform their expectations about new stimuli.

Taken together, our results indicate that the zero-stimulus biases we observed were not due to simple 'response strategies', but rather, were perceptual in origin: participants 'hallucinated' motion in the most frequently presented directions when no stimulus was displayed.

Development of estimation bias when no stimulus present

To investigate how quickly participants' 'no-stimulus' estimation biases developed, we evaluated the probability ratio that individual participants made estimates close to the most frequently presented motion directions, relative to other directions, after every 100 trials (including all responses up to that point; figure 3.6). For participants who had not reported detecting stimuli on any trials where none was presented, this probability ratio was undefined, so these data points were omitted from the plot (e.g. after 100 trials, only 4 participants were included, 11 participants were included after 200 trials, and 12 participants after 300 trials). After only 200 trials of the first session, the median probability ratio (p_{rel}) was significantly greater than 1, indicating that on trials where no stimulus was presented, but where participants reported detecting a stimulus, they were biased to estimate motion in the most frequently presented directions after only 200 trials. Thus, expectations about which motion directions were most likely to occur were learned extremely rapidly, after only a few minutes of task performance.

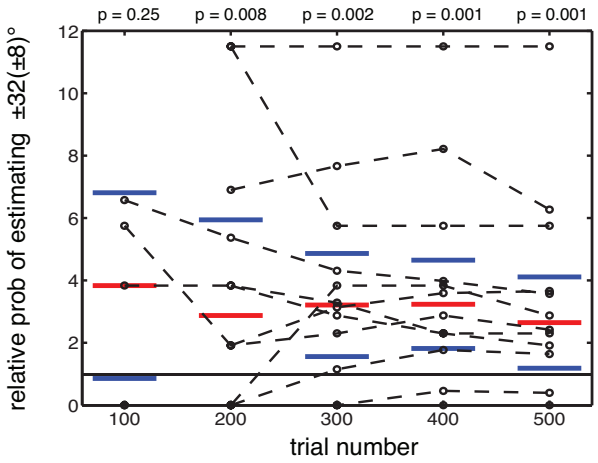


Figure 3.6: Probability ratio that individual participants estimated within 8° from the most frequently presented motion directions (±32°) relative to other 16° windows, for trials where no stimulus was presented, but where they reported detecting a stimulus. This probability ratio is calculated for each participant after every 100 trials (this calculation takes into account data from all trials up to that point; here we show the first 500 trials from the first session only). Median values are indicated by horizontal red lines, 25th and 75th percentiles by horizontal blue lines. Dashed lines correspond to the 'trajectories' of individual participants' 'p_{rel}' values. p-values indicate whether the probability ratio ('p_{rel}') was significantly different from 1 at each point in time.

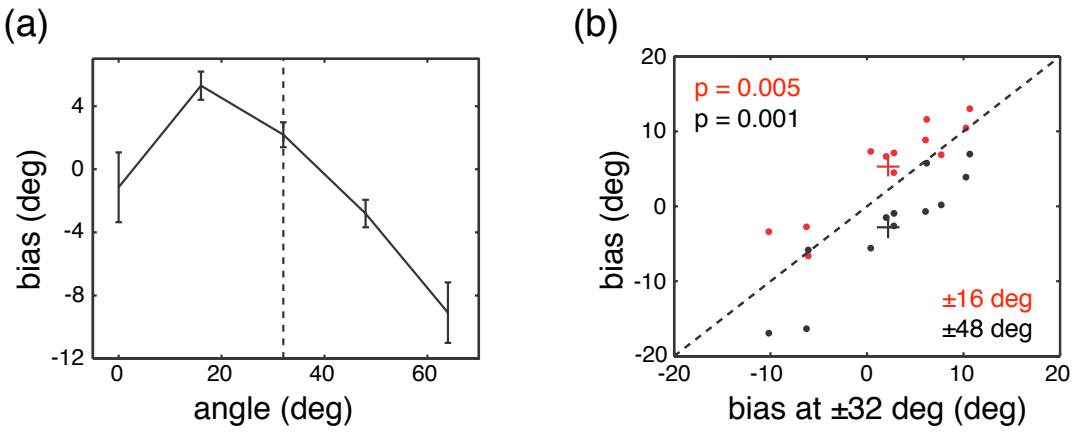


Figure 3.7: Effect of expectations on estimation biases. (a) Participants' mean estimation bias is plotted against presented motion direction. Data points from either side of the central motion direction have been averaged together, so that the furthest left point corresponds to the central motion direction, and the vertical dashed line corresponds to data taken from the two most frequently presented motion directions (±32°). Results are averaged over all participants and error bars represent within-subject standard error. (b) The estimation bias for stimuli moving at ±48° (black) and ±16° (red) from the central motion direction, plotted against the estimation bias at ±32°, for each participant. Again, data from stimuli moving to both sides of the central motion direction has been averaged together, with the sign of the bias for stimuli moving anti-clockwise from the central motion direction (i.e. -48°, -32° and -16°) reversed before averaging. The red and black crosses mark the population mean of both distributions, with the length of the lines on the crosses equal to the standard error.

3.2.3 Estimates of motion direction when stimulus present

We next asked whether participants' learned expectations would bias their perception of real motion stimuli. Figure 3.7a shows the population averaged estimation bias, plotted against motion direction. In this plot, data points corresponding to presented stimuli moving to either side of the central motion direction have been averaged together (making sure to reverse the sign of the estimation bias when the presented stimuli was anti-clockwise from the central motion direction before averaging; see figure A.1 for an alternative version of this plot without averaging across the central motion direction). The plotted curve has a negative slope around +32°, which itself was unbiased. This indicates that estimations were attractively biased towards stimuli moving at +32° (and by symmetry, also to motion at -32°). Estimates of the central motion direction were unbiased, while estimates at +16° were positively biased, away from the centre and towards stimuli moving at +32° (again, by symmetry, stimuli moving at -16° were biased away from the centre, towards stimuli moving at -32°). Note that the apparent

asymmetry in figure 3.2.3a is expected, due to the fact that the data points at 0° and 64° are not equivalent: 0° lies midway between the two most frequently presented directions, while +64° is on the edge of the distribution of presented motion directions (see figure 3.1). Overall, there was a significant effect of motion direction on the estimation bias ($p < 0.001$, 3-way within-subjects ANOVA).

We wanted to quantify the extent to which individual participants' estimates were biased towards the most frequently presented motion directions. For a participant whose estimates were attractively biased towards stimuli moving at +32°, we would expect estimates of stimuli moving at +48° and +16° to be positively and negatively biased respectively, compared to their estimation bias for stimuli moving at +32° (and by symmetry, we would also expect the converse to hold for stimuli moving anti-clockwise from the central direction: for a participant whose estimates were attractively biased towards stimuli moving at -32°, we would expect estimates at -48° and -16° to be negatively biased and positively biased respectively, compared to their estimation bias for stimuli moving at -32°). Figure 3.7b plots individual participants' estimation bias for stimuli moving at ±48° and ±16° versus their estimation bias at ±32° (plotted in black and red respectively). Note that, as with figure 3.7a, we averaged data from motion directions moving to either side of the central motion directions in this plot, making sure to reverse the sign of the bias for stimuli moving anti-clockwise from the central motion direction. After doing this, the computed estimation biases at ±48° and ±16° were significantly smaller and larger respectively than the bias at ±32° ($p = 0.005$ and $p = 0.001$ respectively, signed rank test). This indicates that on average, participants were biased to estimate stimuli as moving in directions that were closer to the most frequently presented motion directions (±32°) than they actually were.

Stimuli in-between ±32° were expected to be biased by both frequently presented directions and thus we expected that these directions should yield larger standard deviations in estimated angles than those outside of this range. Figure 3.8 plots the population-averaged estimation standard deviation versus stimulus motion direction. Again, for this plot, data points from either side of the central motion direction have been averaged together. The estimation standard deviation was greatest for the central motion direction at 0°, and smallest for motion directions that were closer to the most frequently presented directions (±16°, ±32° and ±48°).There was a significant effect of stimulus motion direction on the estimation standard deviation ($p<0.001$, 3-way within-subjects ANOVA).

If participants' learned to expect stimuli moving at $\pm 32°$ we might expect smallest estimation standard deviation for stimuli moving in this direction, when their expectations agree with their received sensory evidence, leading to a reduction in their uncertainty about the presented motion direction. However, we found smallest standard estimation standard deviation for stimuli moving at $\pm 48°$. A potential explanation for this could be that participants did not

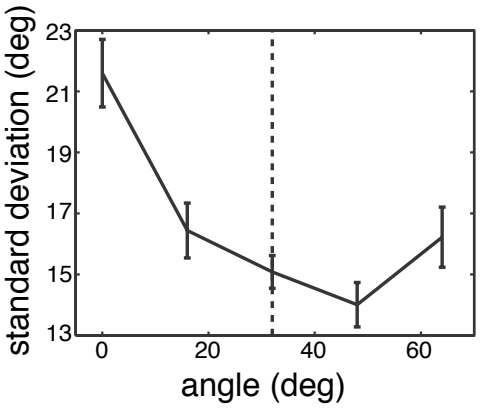


Figure 3.8: Effect of expectations on the standard deviation of estimations. The standard deviation in participants' estimation distributions is plotted against presented motion direction. Data points from either side of the central motion direction have been averaged together, so that the furthest left point corresponds to the central motion direction, and the vertical dashed line corresponds to data taken from the two most frequently presented motion directions (±32°). Results are averaged over all participants and error bars represent within-subject standard error.

learn to expect stimuli moving at exactly $\pm 32^\circ$, but rather were biased to expect stimuli moving towards some other direction, between $\pm 32^\circ$ and $\pm 48^\circ$. This could also account for the small positive estimation bias that we observed for stimuli moving at $\pm 32^\circ$ (figure 3.7 a). However, it does not seem to be reflected in participants' no-stimulus response distributions, whose peak lies at $\pm 32^\circ$.

3.2.3.1 Effect of high contrast stimulus on estimate of subsequent stimulus motion direction

We asked whether a presented high contrast stimulus influenced subjects' perception of a subsequently presented low-contrast stimulus. To test whether this was the case, we analyzed subjects' estimation responses for the subset of trials that directly followed a high contrast stimulus. We quantified how subjects' estimation bias varied as a function of the difference between the direction of the presented stimulus and the previous high contrast stimulus, using Spearman's rank correlation coefficient.

We found that a high contrast stimulus had no systematic effect on subjects' estimation bias for the following stimulus. 6 subjects exhibited a non-significant positive correlation coefficient (equivalent to a repulsive bias away from the high contrast stimulus), and 5 subjects exhibited a non-significant negative correlation coefficient (equivalent to a repulsive bias away from the high contrast stimulus). Only 1 out of 12 subjects exhibited a significant negative correlation coefficient ($r = -0.53$, $p < 0.001$), signifying an attractive estimation bias towards

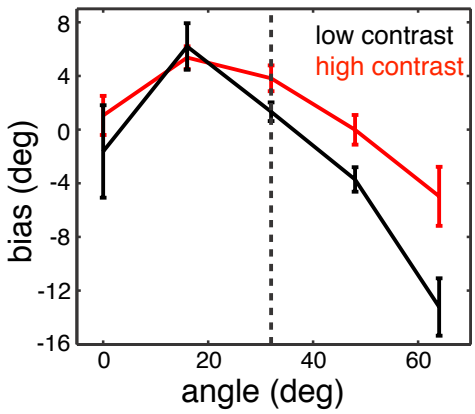


Figure 3.9: Estimation bias at different contrasts levels. Participants' estimation bias for the higher contrast trials (red) and lower contrast trials (black) are plotted against presented motion direction. Data points from either side of the central motion direction have been averaged together, so that the furthest left point corresponds to the central motion direction, and the vertical dashed line corresponds to data taken from the two most frequently presented motion directions (±32°). Results are averaged over all participants, and error bars represent the within-subjects standard error.

the previously presented high contrast stimulus.

Estimation biases at different contrasts

To investigate how stimulus contrast affected participants' estimation behaviour, we first fitted psychometric curves to their detection responses:

$$p_{detect}(c) = \gamma + F(c)(1-\gamma), \tag{3.4}$$

where $p_{detect}(c)$ represents the probability that a participant detected a stimulus presented at a contrast c, γ is a constant representing the probability that a participant reported detecting a stimulus when none was displayed (the 'guess rate'), and $F(c)$ is a cumulative normal distribution (specified by two parameters; the mean and standard deviation of the corresponding normal distribution). We fitted this function to each participant's detection response data, setting γ equal to the fraction of trials where they reported detecting a stimulus when none was presented, and fitting the two parameters of the cumulative normal distribution ($F(c)$) to the data using a simplex algorithm (the Matlab function, 'fminsearch') that maximized their likelihood.

From the psychometric curves obtained for each participant, we selected a 'threshold contrast' c_{trhesh} for each participant, such that $F(c_{thresh}) = 0.75$. We then divided participants' estimation responses into two subsets: trials where the stimulus contrast was greater than c_{thresh},

(referred to as 'high contrast trials') and trials where the stimulus contrast was less than c_{thresh} (referred to as 'low contrast trials'). The population averaged mean luminance for the 'low' and 'high' contrast trials were 0.49±0.02cd/m^2 and 0.61±0.02cd/m^2 above background luminance respectively.

Figure 3.9 plots participants' estimation biases separately for 'low contrast trials' (black) and 'high contrast trials' (red), versus the presented motion direction. Both curves exhibit a qualitatively similar shape: at both contrast levels, estimations of motion stimuli far away from the central motion direction (±64°) were biased towards the central motion direction. This bias reversed close to the central motion direction, so that for both contrast levels, estimations of motion stimuli presented at ±16° were biased away from the central motion direction, and towards the most frequently presented motion directions (±32°).

Importantly however, the magnitude of the estimation biases for stimuli moving far away from the central motion direction (±48° and ±64°) was much larger with the lower contrast stimuli than with the higher contrast stimuli. Overall there was a significant interaction between the effects of the two contrast levels and motion direction on the estimation bias ($p < 0.001$, 3-way within-subjects ANOVA).

There are two surprising features of figure 3.9. First, the bias at $\pm 16^\circ$ is the same for both high and low contrast stimuli. Naively, we might expect a larger bias for the low contrast stimuli, as participants' sensory uncertainty would be larger, causing them to be more strongly influenced by their expectations. However, this result could potentially be explained if, in the low contrast condition, participants' sensory uncertainty was large enough for their estimates of motion direction to be influenced by their expectation for stimuli moving either clockwise *or* anti-clockwise of the central motion direction. In this case, we would expect a smaller estimation bias for stimuli moving at $\pm 16^\circ$, than if participants' estimates were just influenced by their expectation for stimuli moving in the closest expected motion direction.

Second, for high contrast stimuli, there is a positive bias away from the central motion direction. As discussed earlier, this bias could reflect the fact that participants' expect stimuli moving further slightly further away from the central motion direction than $\pm 32^\circ$. However, if this is the case, it is unclear why a similar positive bias is not observed for the low contrast stimuli, unless participants' learn different expectations for different stimulus contrasts.

In general, the estimation standard deviation was significantly larger at the lower contrast level than at the higher contrast level (an average value of 17.8±1.7° at the higher contrast level versus 14.4±1.3° at the lower contrast level; $p = 0.017$, 3-way within-subjects ANOVA). However, there was no significant interaction between the effects of contrast level and presented motion direction on the estimation standard deviation ($p = 0.10$, 3-way within-subjects ANOVA).

These results are qualitatively consistent with what we would expect if participants be-

haved as ideal Bayesian observers. At decreased stimulus contrast the width of participants' sensory likelihood function should increase, with a corresponding increase in their estimation standard deviation. At the same time, participants' estimates of motion direction should be more strongly influenced by their expectations, leading to stronger biases towards the most frequently presented motion directions, as we observed in our experimental data.

3.2.4 Detection performance and reaction time

We investigated how participants' expectations influenced their performance in the detection task. To do this, we measured the fraction of trials where participants both detected stimuli and clicked on the mouse during stimulus presentation, as a function of motion direction (figure 3.10a). Participants were significantly more likely to detect stimuli moving in the most frequently presented motion directions (71.5±2.5% detected at ±32° versus 64.2±2.5% detected over all other motion directions; $p<0.001$ signed-rank test; figure 3.10b). Overall, there was a significant effect of motion direction on the fraction detected ($p = 0.002$, 3-way within-subjects ANOVA).

Another measure that could reflect how easily participants detected stimuli was their reaction time in clicking the mouse during stimulus presentation. Thus, we measured participants' reaction times in the estimation task as a function of stimulus motion direction (figure 3.10c). For trials where they detected a stimulus, participants' reaction time was significantly reduced for the most frequently presented motion directions, relative to other motion directions (1924±86ms at ±32° versus 1991±85ms over all other motion directions; $p < 0.001$, signed rank test; figure 3.10d). Overall, there was a significant effect of motion direction on participants' reaction time ($p = 0.003$, 3-way within-subjects ANOVA).

3.3 Modelling

To understand the nature of the biases in motion direction estimation that we observed, we tested alternative models of how participants' expectations could be combined with the presented stimulus to produce the observed response distributions. Two classes of models were considered. The first class of model assumes that participants developed response strategies that were unrelated to perceptual changes (section 3.3.1). The second class of model assumes that participants solved the task using a Bayesian strategy, combining a learned prior of the stimulus statistics (the expectation) with their sensory evidence (the actual stimulus) using Bayes' rule (section 3.3.2).

In section 3.3.1 & 3.3.2, we just consider participants' behaviour in the estimation task. Later, in section 3.3.5, we consider their behaviour in both the estimation and the detection task. Finally, we use our Bayesian model to predict participants' estimation responses on trials

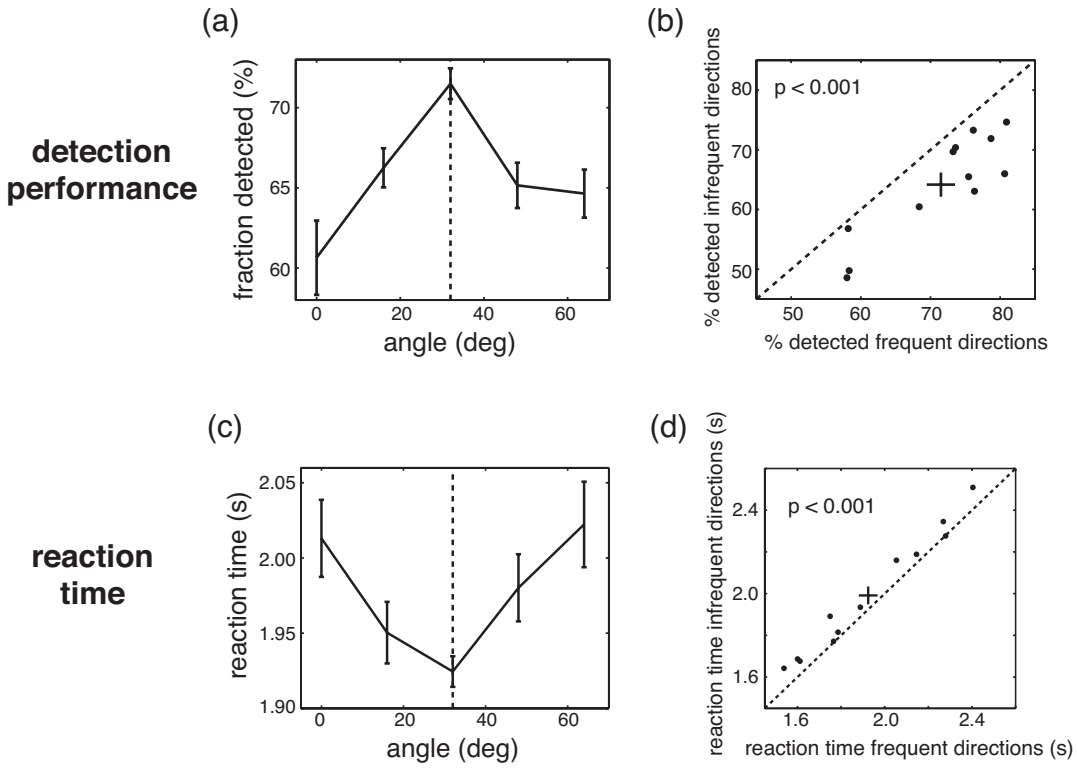


Figure 3.10: Effect of expectations on detection performance and reaction times. (a) The fraction of trials where participants correctly detected a motion stimulus, plotted against the presented motion direction. (b) The fraction of trials where participants correctly detected a stimulus, averaged over all presented motion directions except for ±32°, plotted against the fraction of trials where participants correctly detected a stimulus moving at ±32°, for each participant. (c) Time taken for participants to click on the mouse during stimulus presentation, measured from the initial presentation time. (d) Individual average reaction time for stimuli moving at ±32°, plotted against the reaction time over all other motion directions. In panels (a) & (c), data points from either side of the central motion direction have been averaged together, so that the furthest left point corresponds to the central motion direction, and the vertical dashed line corresponds to the most frequently presented motion directions (±32°). Results are averaged over all participants and error bars represent within-subject standard error. In panels (b) & (d), the black cross marks the population mean, with the length of the lines on the cross equal to the standard error.

where no stimulus was presented.

The small number of trials for each staircased contrast level, as well as the large degree of overlap between the luminance levels for each staircase (see figure 3.3), meant that it was difficult to constrain the model parameters for each contrast level separately. Therefore, for all our modelling work we analyzed trials from both staircased contrast levels together. Thus, future modelling work will be required to quantify how subjects' expectations alter their estimation behaviour at different stimulus contrasts.

3.3.1 Multiple strategy 'response-bias' models

The first class of model assumes that participants' behaviour can be attributed to a 'response bias'. The key assumption of these models is that participants follow different strategies on different trials: for example, by making an unbiased estimate of motion direction on a fraction of the trials, and by estimating one of the most frequently presented motion directions on other trials.

The first model ('*ADD1*') assumes that when participants were unsure about which motion direction they had perceived, they made an estimate that was close to one of the two most frequently presented motion directions.

On each trial, participants make a 'sensory observation' of the stimulus motion direction, θ_{obs}, that depends on their received sensory input. Given a stimulus moving in a direction θ, the probability of observing motion direction θ_{obs} is described by a von Mises (circular normal) distribution centred on the actual stimulus direction (θ) and with width determined by $1/\kappa_l$:

$$p_l(\theta_{obs}|\theta) = \mathcal{V}(\theta_{obs};\theta,\kappa_l) \tag{3.5}$$

On most trials, participants are assumed to make a perceptual estimate of the stimulus motion direction (θ_{perc}) that is based entirely on their sensory observation: $\theta_{perc} = \theta_{obs}$. However, on a certain proportion of trials, when participants are uncertain about whether a stimulus is present or not, they resort to their 'expectations', making a perceptual estimate that is sampled from a distribution of expected motion directions: $p_{exp}(\theta)$. For simplicity, we parameterize this distribution as the sum of two circular normal distributions, each with width determined by $1/\kappa_{exp}$, and centred on motion directions $-\theta_{exp}$ and θ_{exp} respectively:

$$p_{exp}(\theta) = \frac{1}{2}\left(\mathcal{V}(\theta;-\theta_{exp},\kappa_{exp}) + \mathcal{V}(\theta;\theta_{exp},\kappa_{exp})\right) \tag{3.6}$$

We accommodate for the noise associated with moving the estimation bar to indicate which direction the stimulus is moving in, as well as allowing for a fraction of trials 'α', where participants make estimates that are completely random. Thus, participants' estimation responses θ_{est} are related to the perceptual estimate θ_{perc} via:

$$p(\theta_{est}|\theta_{perc}) = (1-\alpha)\,\mathcal{V}(\theta_{est};\theta_{perc},\kappa_m) + \frac{\alpha}{2\pi} \tag{3.7}$$

Together equations 3.5, 3.6 & 3.7 determine the distribution of estimations:

$$p(\theta_{est}|\theta) = (1-\alpha)\Big[(1-a(\theta))\,p_l(\theta_{est}|\theta) + a(\theta)\,p_{exp}(\theta_{est})\Big] \star \mathcal{V}(\theta_{est};0,\kappa_m) + \frac{\alpha}{2\pi} \tag{3.8}$$

where '$\star$' denotes a convolution and '$a(\theta)$' determines the proportion of trials where participants sample from the 'expected' distribution, $p_{exp}(\theta)$. Free parameters that were fitted to the estimation data for each participant were the centre and width of $p_{exp}(\theta)$ (determined by θ_{exp} & $1/\kappa_{exp}$ respectively), the width of $p_l(\theta_{obs}|\theta)$ (determined by $1/\kappa_l$), the fraction of trials where they made estimates by sampling from the $p_{exp}(\theta)$ ($a(\theta)$), the 'motor' noise in their estimation responses (determined by $1/\kappa_m$) and the fraction of trials where they made random estimates (α).

The second 'response-bias' model ('*ADD2*') assumes a more complex strategy; that, when participants are unsure about the stimulus motion direction, they make estimates that are preferentially sampled from different portions of the 'expected' distribution, depending on the presented motion direction (θ).

As before, on a single trial, participants are assumed to make estimates that are either equal to their sensory observation θ_{obs}, or sampled from a distribution of expected motion directions. However, instead of sampling from a single distribution of expected motion directions (as was the case for the *ADD*1 model) participants are assumed to sample either from $p_{anti-clockwise}(\theta) = \mathcal{V}(\theta;-\theta_{exp},\kappa_{exp})$ or from $p_{clockwise}(\theta) = \mathcal{V}(\theta;\theta_{exp},\kappa_{exp})$, with a probability that depends on the presented motion direction. For example, on a single trial, a participant might be aware that the stimulus is moving 'anti-clockwise from centre', and they would be more likely to sample from $p_{anti-clockwise}(\theta)$, than from $p_{clockwise}(\theta)$. This more complex response strategy results in a distribution of estimation responses given by:

$$p(\theta_{est}|\theta) = (1-\alpha)\Big[(1-a(\theta)-b(\theta))\,p_l(\theta_{est}|\theta) + a(\theta)\,p_{anti-clockwise}(\theta_{est}) + b(\theta)\,p_{clockwise}(\theta_{est})\Big] \star \mathcal{V}(\theta_{est};0,\kappa_m) + \frac{\alpha}{2\pi} \tag{3.9}$$

where '$a(\theta)$' and '$b(\theta)$' are additional free parameters that determine the proportion of trials where participants sample from each distribution.

Model variants

We considered several different variants of the *ADD*1 and *ADD*2 models, which made different assumptions about how participants' response strategy varied with the presented motion direction, and the way in which they sampled from $p_{exp}(\theta)$. The model variants are as follows:

- $ADD1$: 9 free parameters. The fraction of trials where participants sample from $p_{exp}(\theta)$ (parameterized by $a(\theta)$) is assumed to vary for each presented motion direction (we assume symmetry about the central motion direction, so that $a(\theta) = a(-\theta)$).

- $ADD1_{reduced}$: 6 free parameters. The fraction of trials where participants sample from $p_{exp}(\theta)$ is assumed to vary linearly with the presented motion direction, as $a(\theta) = m|\theta| + c$. A linear fit was chosen as it represents the 'simplest' possible parameterization for how subjects response strategy could have varied with the presented motion direction.

- $ADD1_{mode}$: 8 free parameters. On trials where participants are unsure of the stimulus motion direction, we assume they make perceptual estimates that are equal to the *mode* of the 'expected' distribution. This assumption is equivalent to setting '$1/\kappa_{exp}$' to zero (i.e. in the limit where $\kappa_{exp} \to \infty$).

- $ADD1_{reduced,mode}$: 5 free parameters. A combination of the assumptions used in the $ADD1_{reduced}$ and $ADD1_{mode}$ models, so that $1/\kappa_{exp} = 0$, and $a(\theta) = m|\theta| + c$

- $ADD2$: 14 free parameters. The fraction of trials where participants' sample from $p_{anti-clockwise}(\theta)$ and $p_{clockwise}(\theta)$ (parameterized by $a(\theta)$ and $b(\theta)$) is assumed to vary for each presented motion direction.

- $ADD2_{reduced}$: 8 free parameters. The fraction of trials where participants' sample from $p_{exp}(\theta)$ is assumed to vary linearly with the presented motion direction: $a(\theta) = m\theta + c$, $b(\theta) = n\theta + d$

- $ADD2_{mode}$: 13 free parameters. On trials where participants are unsure of the stimulus motion direction, we assume they make perceptual estimates that are equal to the *mode* of the 'expected' distribution (i.e. $1/\kappa_{exp} = 0$)

- $ADD2_{reduced,mode}$: 7 free parameters. A combination of the assumptions used in the $ADD2_{reduced}$ and $ADD2_{mode}$ models, so that $1/\kappa_{exp} = 0$, $a(\theta) = m\theta + c$, and $b(\theta) = n\theta + d$

3.3.2 Bayesian model

The second class of model assumes that participants combine a learned prior of the stimulus directions with received sensory information using Bayes' rule. Unlike the previous models, which assume that on a single trial participants rely either on their sensory observations or on their expectations, the Bayesian models assume that on each trial participants combine their sensory evidence and expectations to estimate the stimulus motion direction. A schematic of this model class is shown in figure 3.11.

As before, we assume that on a single trial, participants make noisy sensory observations of the stimulus motion direction (θ_{obs}), with probability, $p_l(\theta_{obs}|\theta) = V(\theta_{obs}; \theta, \kappa_l)$. Participants

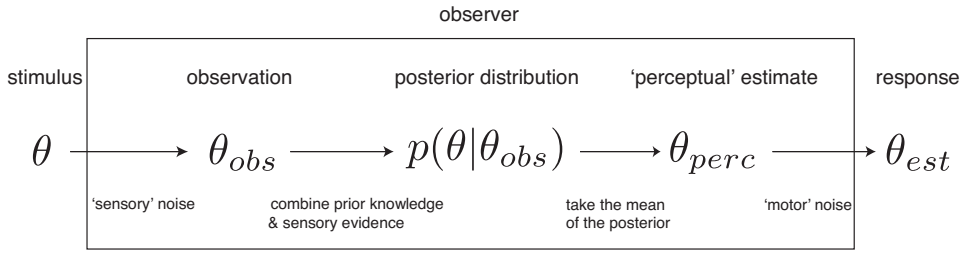


Figure 3.11: Bayesian model of estimation responses. The posterior distribution of possible stimulus motion directions is constructed by combining prior knowledge about likely motion directions (the expectation) with the available sensory evidence (based on a noisy observation, θ_{obs}) probabilistically. A perceptual estimate is made by taking the mean of the posterior distribution. This posterior distribution is used to make a perceptual estimate (θ_{perc}). Additional 'motor noise' is added to this perceptual estimate to produce the final estimation response (θ_{est}).

are assumed to learn a prior distribution over motion directions ($p_{exp}(\theta)$; parameterized as in equation 3.6). We assume that this learned prior is combined with their received sensory evidence using Bayes' rule:

$$p(\theta|\theta_{obs}) \propto p_{exp}(\theta)\, p_l(\theta_{obs}|\theta)\,. \tag{3.10}$$

Participants' are assumed to make a perceptual estimate θ_{perc} that is equal to the mean of the posterior distribution:

$$\theta_{perc} = \frac{1}{Z}\int \theta p_{exp}(\theta)\, p_l(\theta_{obs}|\theta)\, d\theta, \tag{3.11}$$

where $Z = \int p(\theta_{exp})\, p_l(\theta_{obs}|\theta)\, d\theta$. Qualitatively similar results were obtained when participants were assumed to make perceptual estimates that were equal to the maximum of the posterior distribution.

We accounted for the 'motor noise' associated with making an estimation response in the same way as for the previous 'response-bias' models (equation 3.7). For the Bayesian model, free parameters that were fitted to the estimation data for each participant were the centre and width of the prior (determined by θ_{exp} & $1/\kappa_{exp}$ respectively), the width of the sensory likelihood (determined by $1/\kappa_l$), the 'motor noise' (determined by $1/\kappa_m$) and the fraction of trials where participants' made random estimates (α).

Model variants

We considered two variants of the Bayesian model:

- '*BAYES_var*': 9 free parameters. The width of the likelihood function (κ_l) could vary with the presented stimulus motion direction.
- '*BAYES*'. 5 free parameters. κ_l was set to be the same for all presented motion directions.

3.3.3 Fitting the model parameters

As the high contrast motion stimuli were clearly visible (correctly detected on 97±0.3% of trials), we assumed that the perceptual uncertainty was close to zero for these stimuli ($1/\kappa_l \approx 0$). In this case, all of the described models predict that participants' estimation responses should be distributed according to:

$$p(\theta_{est}|\theta) = (1-\alpha)\,\mathcal{V}(\theta_{est};\theta,\kappa_m) + \frac{\alpha}{2\pi}. \tag{3.12}$$

We fitted the free parameters in this expression (κ_m and α) to participants' estimation responses with the high contrast stimuli (by maximizing the log-likelihood of the model parameters). Estimates of participants' motor noise obtained with the high contrast stimuli were used to fit their estimation responses with the low contrast stimuli (under the assumption that the 'motor-noise',$1/\kappa_m$, for each participant is independent of stimulus contrast).

As all three models only described participants responses in the estimation task, ignoring their detection responses, we just looked at data where participants' correctly detected a presented motion stimulus (see section 3.3.5 for a Bayesian model of the detection task). For each model, we fitted the estimation responses for individual participants, by maximizing the log-likelihood of the model parameters $\mathcal{M}$:

$$\mathcal{M} = \arg\max_{\mathcal{M}} \sum_{i=1}^{N} \log p\left(\theta^i_{est}|\theta^i,\mathcal{M}\right), \tag{3.13}$$

where θ^i_{est} & θ^i_{data} denote the presented motion direction and estimation response on the i^{th} trial, respectively. The likelihood function was maximized using a simplex algorithm (the Matlab function 'fminsearch'). We were concerned that for some participants, our model fits might converge to local rather than global maxima. To reduce this possibility, we ran the model fits with a range of initial values for κ_l and κ_{exp} ($\kappa_l^{-1/2}$ and $\kappa_{exp}^{-1/2}$ were varied independently in 2° increments, between 1° and 21°), selecting the model fit that produced the highest value for the log-likelihood. The results obtained were also found to be robust to changes in all of the other initial parameter values.

3.3.4 Model comparison

3.3.4.1 Statistical comparison of models

We assessed how well each of the models were able to account for participants' estimation distributions using a metric called the 'Bayesian information criterion' (*BIC*):

$$BIC = -2\log(\mathcal{L}) + k\log(n), \tag{3.14}$$

where, '$\mathcal{L}$' denotes the maximized value of the likelihood for the estimated model, 'k' is the number of model parameters and 'n' is the number of data points. The first term of this expression describes how well the model is able to fit the data, while the second term represents a

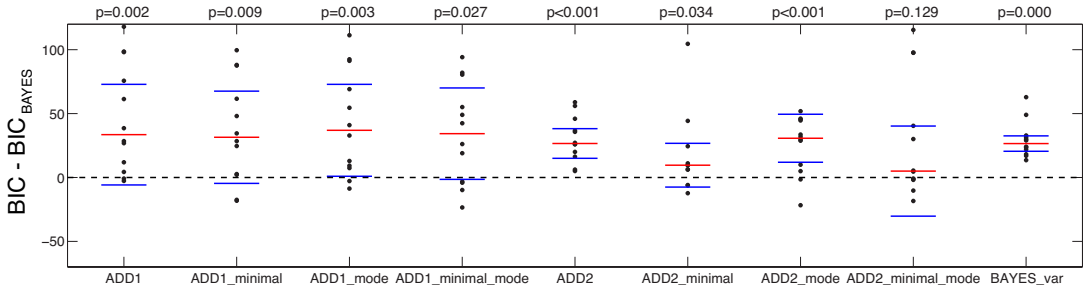


Figure 3.12: Model comparison. The Bayesian information criterion (*BIC*) evaluated with each model, subtracted by the *BIC* evaluated with the *BAYES* model, plotted separately for each participant. Median values are indicated by horizontal red lines, 25th and 75th percentiles by horizontal blue lines. A *BIC* value greater than zero indicates that the *BAYES* model provided a better description of the estimation data. p-values indicate whether the *BIC* value for each model is significantly different than the *BIC* value for the *BAYES* model (signed rank test).

penalty for including too much complexity in the model. Given two model fits, the model with the lower *BIC* value is usually the one to be preferred (Schwarz, 1978).

Figure 3.12 plots, for each participant, the *BIC* obtained with each model, subtracted by the *BIC* obtained with the *BAYES* model. The *BIC* values were significantly higher for all of variants of the $ADD1$ model than for the $BAYES$ model (p=0.002, p=0.009, p=0.003, p=0.027, for the $ADD1$, $ADD1_{minimal}$, $ADD1_{mode}$ and $ADD1_{minimal,mode}$ models respectively; signed rank test). Likewise, the *BIC* value was significantly higher for the $BAYES_{full}$ model than for the $BAYES$ model (p<0.001; signed rank test). For 3 variants of the $ADD2$ model the *BIC* value was significantly higher than for the $BAYES$ model (p<0.001, p=0.034, p=0.005 for the $ADD2$, $ADD2_{minimal}$, and $ADD2_{mode}$ models respectively; signed rank test). However, there was no significant difference between the *BIC* values obtained with the $ADD2_{minimal,mode}$ and the $BAYES$ model (p=0.129; signed rank test).

To investigate the extent to which differences in the observed *BIC* values depended on the '$k\log(n)$' penalty associated with the number of model parameters, we plotted the difference between the log-likelihood for the $BAYES$ model and the log-likelihood for each of the other models ($\log(\mathcal{L}_{BAYES}) - \log(\mathcal{L})$; figure 3.13). For all of the models except for the $ADD2_{mode}$ and the $ADD2_{minimal,model}$ models, the log-likelihood was not significantly different from the *BAYES* model. The log-likelihood for the $ADD2_{minimal}$ and $ADD2_{minimal,mode}$ model was significantly greater than the $BAYES$ model ($p < 0.001$ for both the $ADD2_{minimal}$ and the $ADD2_{minimal,mode}$ models; signed rank test). Thus, while the $BAYES$ model exhibited smaller *BIC* values than the response-strategy models, this difference was mainly due to the fewer number of model parameters in the $BAYES$ model, rather than the quality of the model fit.

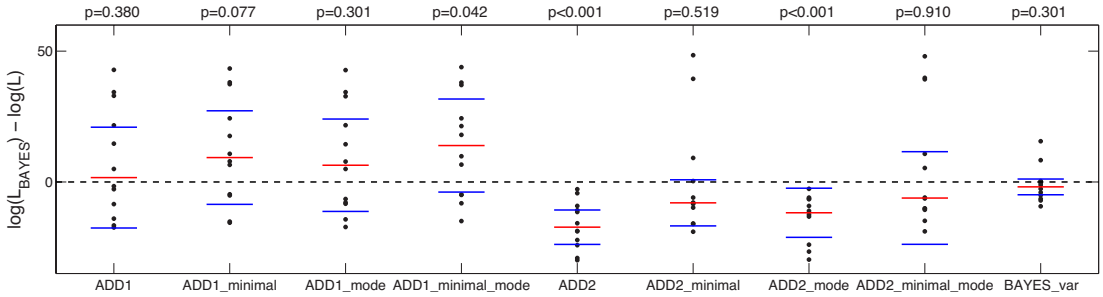


Figure 3.13: Log-likelihood of the BAYES model, subtracted by the log-likelihood of each of the other models. Each point corresponds to one participant. Median values are indicated by horizontal red lines, 25th and 75th percentiles by horizontal blue lines. p-values indicate whether the log-likelihood for each model is significantly different from the the log-likelihood for the *BAYES* model (signed rank test).

3.3.4.2 Fits of estimation bias and standard deviation

To achieve a qualitative understanding of how the models were able to fit participants' estimation responses, we analyzed the average estimation bias and standard deviation obtained with each of the models. In our previous analysis of the experimental data, we parameterized participants' estimation distributions as the sum of a circular normal distribution and a 'flat' background probability (the proportion of trials where they were assumed to make random estimations). Participants' estimation means and standard deviations were then taken as the centre and width of the fitted circular normal distribution respectively. For consistency, we computed biases and standard deviations from the estimation distributions predicted by each model in an identical way.

Figure 3.14 shows the estimation bias and standard deviation obtained with the *BAYES* model, plotted alongside the experimentally measured estimation bias and standard deviation. The *BAYES* model is able to provide a good fit of participants' estimation biases, with an attractive estimation bias towards stimuli moving $\pm 32^\circ$ from the central motion direction. While the predicted estimation standard deviation is larger than was observed experimentally, the *BAYES* model predicts that the estimation standard deviation should vary with the presented motion direction in a qualitatively consistent way to the experimental data, with largest estimation standard deviation for stimuli moving in the central motion direction.

We next consider the estimation bias and standard deviation obtained with the $ADD2_{mode,minimal}$ model, which had the best *BIC* values out of all the 'response-strategy' models, and was the only model whose *BIC* value was not significantly different from the *BAYES* model (figure 3.12). The estimation biases predicted by the $ADD2_{mode,minimal}$ vary with the presented motion direction in a qualitatively manner to the experimental data (3.15a). The $ADD2_{mode,minimal}$ model predicts a large estimation biases for stimuli moving in the central motion direction,

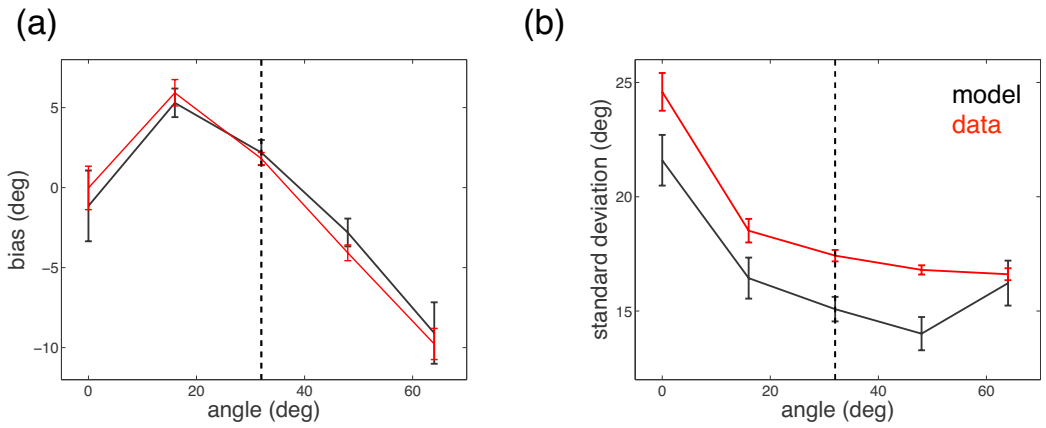


Figure 3.14: Predicted biases (a) and standard deviations (b) for the *BAYES* model (red), plotted alongside the experimental data (black). In both plots, data points from either side of the central motion direction have been averaged together, so that the furthest left point corresponds to the central motion direction, and the vertical dashed line corresponds to the most frequently presented motion directions. Results are averaged over all participants and error bars represent within-subject standard error.

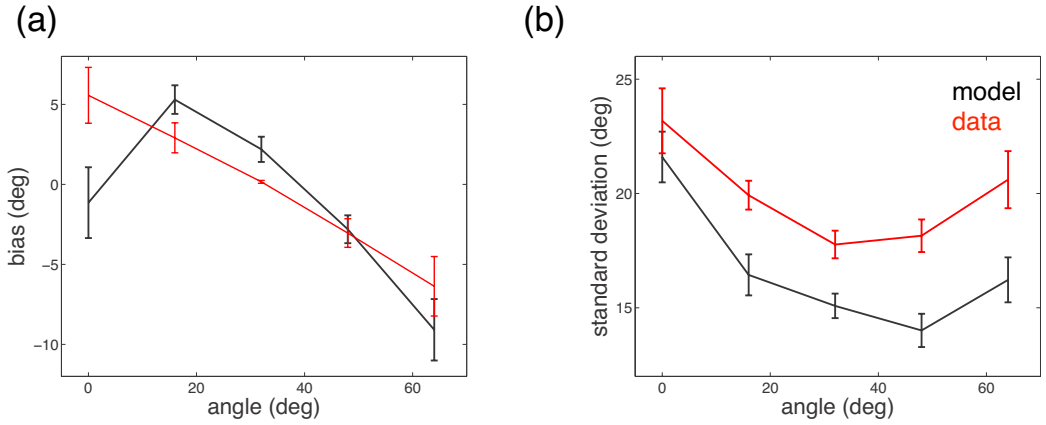


Figure 3.15: Predicted biases (a) and standard deviations (b) for the $ADD2_{minimal\,reduced}$ model (red), plotted alongside the experimental data (black). In both plots, data points from either side of the central motion direction have been averaged together, so that the furthest left point corresponds to the central motion direction, and the vertical dashed line corresponds to the most frequently presented motion directions. Results are averaged over all participants and error bars represent within-subject standard error.

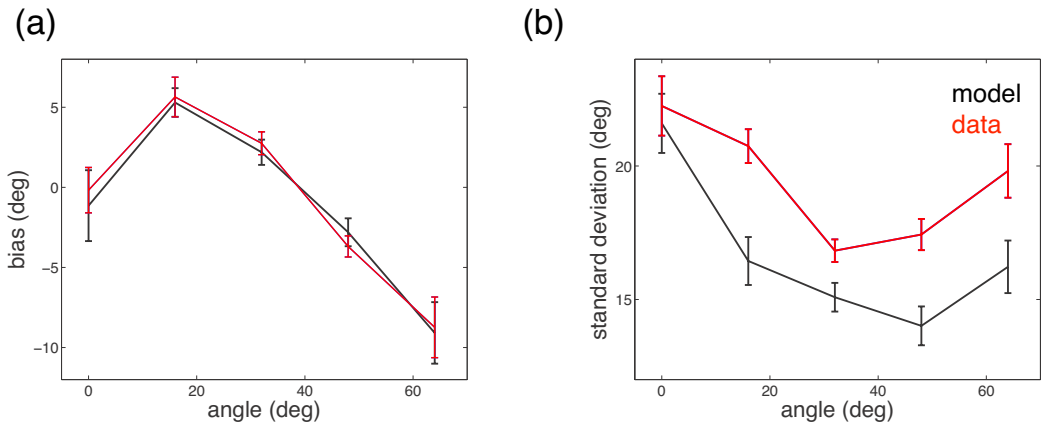


Figure 3.16: Predicted biases (a) and standard deviations (b) for the $ADD2_{minimal\,reduced}$ model (red), plotted alongside the experimental data (black).

decreasing linearly with the presented stimulus direction . Thus, while the *BIC* value for the $ADD2_{mode,minimal}$ model is not significantly different from the $BAYES$ model, the $ADD2_{mode,minimal}$ is unable to capture certain key features of participants' estimation behaviour. As with the $BAYES$ model, the estimation standard deviation obtained with the $ADD2_{mode,minimal}$ model is larger than what was observed experimentally (3.15b).

We next consider the estimation bias and standard deviation obtained with the $ADD2_{mode}$ model, in which the fraction of trials where participants made estimates that were equal to either one of the expected motion directions were– free to vary with the presented motion direction (figure 3.16). Unlike the $ADD2_{mode,minimal}$ model, the $ADD2_{mode}$ model was able to provide a very good fit to participants' estimation biases, with an attractive bias towards stimuli moving in $\pm 32^\circ$ (figure 3.16a). As with the other models, the predicted estimation standard deviation was larger than was observed experimentally (figure 3.16b).

Finally we looked at the estimation bias and standard deviation predicted by the $ADD1$ model (figure 3.17). Unlike the $BAYES$ and $ADD2$ models, the $ADD1$ model was unable to fit the repulsive estimation bias away from the central motion direction when stimuli were moving in $\pm 16^\circ$ (figure 3.17a). This can be explained by the fact that for the $ADD1$ model we parameterized the 'expected' distribution of motion directions, $p_{exp}(\theta)$, to be symmetrical around 0°. Thus, even in the extreme case where all responses are sampled from this distribution, there would only be an attractive bias towards the central motion direction. As with the other models, the $ADD1$ model predicted estimation standard deviations that were larger than what was observed experimentally (figure 3.17b).

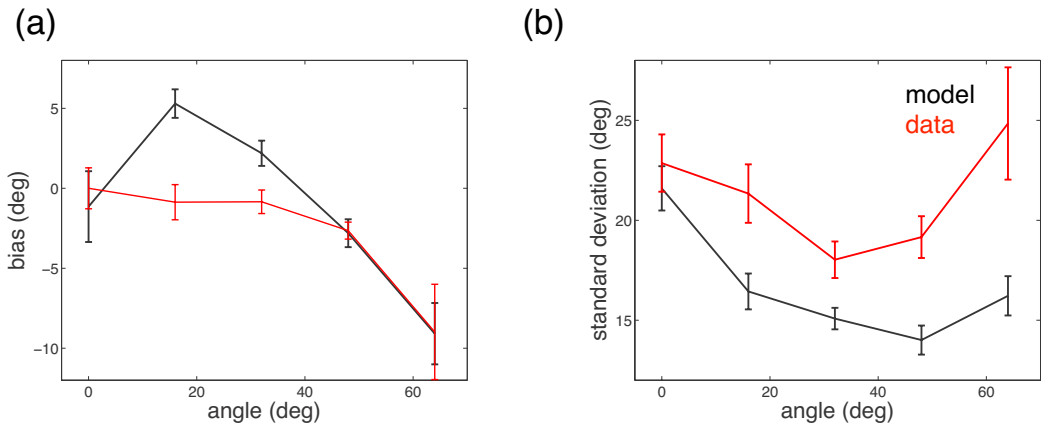


Figure 3.17: Predicted biases (a) and standard deviations (b) for the $ADD1$ model (red), plotted alongside the experimental data (black).

3.3.4.3 Summary of model comparison

The $BAYES$ model exhibited a significantly lower BIC value than all of the competing models apart from the $ADD2_{mode,minimal}$ model, whose BIC value was not significantly different from the $BAYES$ model. The main reason that the $BAYES$ model exhibited lower BIC values was due to the fewer number of parameters; the log-likelihood for the other models was similar, and in some cases better than for the $BAYES$ model.

The $BAYES$ model was able to provide a good qualitative description of how participants' estimation biases and standard deviation varied with the presented motion direction. In contrast, many of the response bias models were unable to reproduce certain key aspects of the data. The 'response strategy' model with the lowest BIC value (the $ADD2_{minimal,mode}$ model), predicted a large estimation bias away from the central motion direction, which was not observed in the data. This discrepancy came about due to the assumptions that the $ADD2_{minimal,mode}$ makes about how participants vary their response strategy depending on the presented motion direction. Without these assumptions (as in the $ADD2$ model), participants' estimation biases can be well fitted by the response strategy model. However, this comes at the expense of an additional 5 parameters, and thus a significantly larger BIC value. The simplest class of response strategy models ($ADD1$) models could not reproduce the repulsive estimation biases away from the central motion direction that we observed experimentally.

3.3.5 Modelling the detection task

We constructed a Bayesian model to describe participants' behaviour in both the estimation and the detection task ($BAYES\_dual$). The motivation for this model was twofold. First, we were concerned that participants' behaviour in the detection task could have altered their behaviour

in the estimation task. Therefore, it was important to check whether our model of participants' behaviour in the estimation task only (*BAYES*) gave consistent results to a model that described their responses in both the estimation and the detection task (*BAYES_dual*). Second, we wanted to investigate whether participants' behaviour in the detection task could be explained within a Bayesian framework.

On a single trial, stimuli move in a direction θ, and can be either present ($s = 1$) or not present ($s = 0$). Participants make sensory observations $\{\theta_{obs}, s_{obs}\}$ with a probability: $p_l(\theta_{obs}, s_{obs}|\theta, s)$. They are assumed to evaluate the posterior probability distribution over s and θ using Bayes' rule:

$$p(\theta, s|\theta_{obs}, s_{obs}) \propto p_l(\theta_{obs}, s_{obs}|\theta, s)\, p_{exp}(\theta, s). \tag{3.15}$$

For simplicity, sensory observations of whether the stimulus is present (s_{obs}) are assumed to be independent of sensory observations of motion direction (θ_{obs}), given the presented stimulus (defined by θ and s), so that the likelihood function can be factorized as follows:

$$p_l(\theta_{obs}, s_{obs}|\theta, s) = p_l(\theta_{obs}|\theta, s)\, p_l(s_{obs}|\theta, s). \tag{3.16}$$

We parameterize the likelihood function over θ_{obs} as:

$$p_l(\theta_{obs}|\theta, s) = \begin{cases} \frac{1}{2\pi} & \text{if } s = 0 \\ \mathcal{V}(\theta_{obs}; \theta, \kappa_l) & \text{if } s = 1. \end{cases} \tag{3.17}$$

Thus, for trials where no stimulus is presented, we assume that participants are equally likely to observe the stimulus to be moving in any direction.

The likelihood function over s is parameterized as:

$$p_l(s_{obs} = \{0, 1\}|\theta, s) = \begin{cases} \{1-c, c\} & \text{if } s = 0 \\ \{1-d, d\} & \text{if } s = 1. \end{cases} \tag{3.18}$$

Our previous Bayesian model of participants' estimation responses did not provide a better fit to the data when κ_l was allowed to vary with the presented motion direction (figure 3.12). Thus, for the *BAYES_dual* model presented here, we assumed that the shape of the likelihood function does not vary with the presented motion direction (i.e. 'κ_l' and 'd' are fixed).

Participants' learned prior ($p_{exp}(\theta, s)$) is parameterized as:

$$p_{exp}(\theta, s) = \begin{cases} \frac{1}{2\pi}(1-b) & \text{if } s = 0 \\ \frac{b}{2}\left(V(-\theta_{exp}, \kappa_{exp}) + V(\theta_{exp}, \kappa_{exp})\right) & \text{if } s = 1, \end{cases} \tag{3.19}$$

where 'b' denotes the prior probability that a stimulus is presented.

Participants are assumed to perform the detection task by taking the maximum of the posterior distribution on each trial (as 's_{perc}' is a discrete binary variable, it does not make

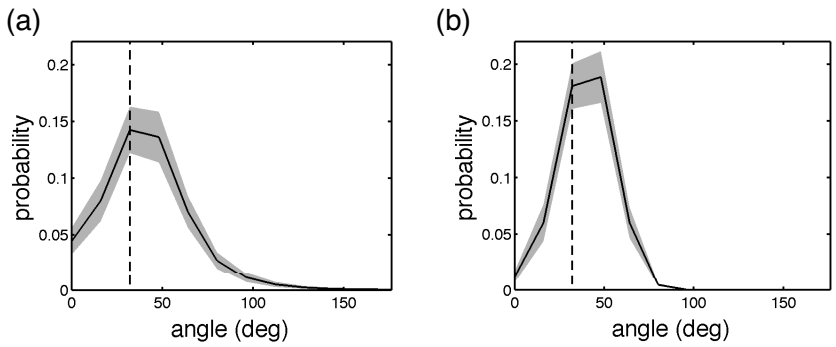


Figure 3.18: Participants' learned prior over the presented motion directions, predicted by the *BAYES* model (a) and the *BAYES_dual* model (b). Data points from either side of the central motion direction have been averaged together in both plots, so that the furthest left data point corresponds to the central motion direction, and the vertical dashed line corresponds to the most frequently presented motion directions (±32°). Results are averaged over all participants and error bars represent within-subject standard error.

sense to estimate the mean of the posterior). Thus, they judge the stimulus to be present if $p(s|\theta_{obs}, s_{obs}) > 0.5$, and absent otherwise/

To be consistent with our previous *BAYES* model, we assume that participants make estimations of motion direction that are equal to mean of the posterior:

$$\begin{aligned} \theta_{perc} &= \int \theta p(\theta|\theta_{obs}, s_{obs})\, d\theta \\ &= \frac{1}{Z} \int \theta \sum_{s=0}^{1} p_l(s_{obs}|s)\, p_l(\theta_{obs}|\theta, s)\, p_{exp}(\theta, s)\, d\theta, \end{aligned} \tag{3.20}$$

where $Z = \sum_{s=0}^{1} \int p_l(s_{obs}|s)\, p_l(\theta_{obs}|\theta, s)\, p_{exp}(\theta, s)\, d\theta$ ensures that the posterior probability–distribution is normalized. As with the *BAYES* model, we obtained qualitatively similar results when we assumed that participants' made estimates that were equal to the maximum of the posterior, instead of the mean. We accounted for the 'motor noise' associated with making an estimation response in the same way as for the previously described models (equation 3.7). In total, the *BAYES_dual* model had 7 free parameters that were fitted to the data for each participant: α, κ_l, c, d, θ_{exp}, κ_{exp}, and b.

Model parameters were fitted to the data using a maximum-likelihood procedure (as in section 3.3.3). Parameters were fitted to both participants' estimation and detection responses on trials where a stimulus was presented. Note that participants' responses on trials where no stimulus was presented were not used to fit the model parameters. Thus, the predicted estimation behaviour on these 'no stimulus' trials could be used to validate the model (see later).

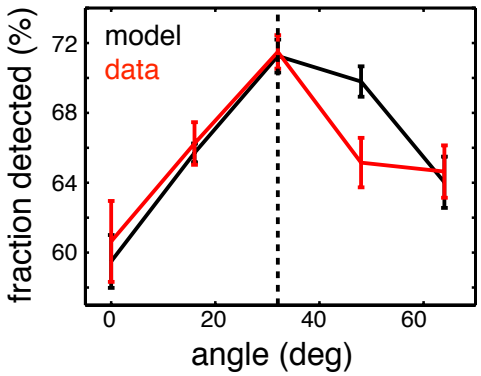


Figure 3.19: Fraction of motion stimuli that were detected, plotted against presented motion direction. Model fit is plotted in black, experimental data is plotted in red. Data points from either side of the central motion direction have been averaged together, so that the furthest left point corresponds to the central motion direction, and the vertical dashed line corresponds to data taken from the two most frequently presented motion directions (±32°). Results are averaged over all participants and error bars represent within-subject standard error.

3.3.5.1 Shape of the prior

Figure 3.18a & b plot the shape of participants' learned prior, required by the *BAYES* model and the *BAYES_dual* model, respectively, to fit the experimental data. The exact shape of the predicted distributions varied between the two models: the *BAYES* model predicted a broader distribution than the *BAYES_dual* model. Indeed, even within each model, there were considerable variations in the location and width of the peaks between individual participants. However, the shape of the population averaged 'learned prior' distributions was qualitatively similar for both models: with a peak lying close to the frequently presented directions (±32°), and falling off at the central motion direction (0°) and to either side of the frequently presented directions (greater than +64° or less than -64°). Notably, both of these distributions had a qualitatively similar shape to the true stimulus distribution (figure 3.1).

3.3.5.2 Fit of participants' detection and estimation responses, when stimulus present

The *BAYES_dual* model provided a reasonable qualitative fit for participants' responses in the detection task (a mean absolute error of 1.50±0.58% detected; figure 3.19). The model exhibited increased detection performance for the most frequently presented motion directions, similar to what was observed experimentally (the model predicted 71.2±1.6% detected at ±32° versus 64.8±1.5% other motion directions compared to experimental observations of 71.5±2.5% detected at ±32°, versus 64.2±2.5% for other motion directions). However, there is some discrepancy between the model predictions and the data at ±48°, where the model predicts a larger

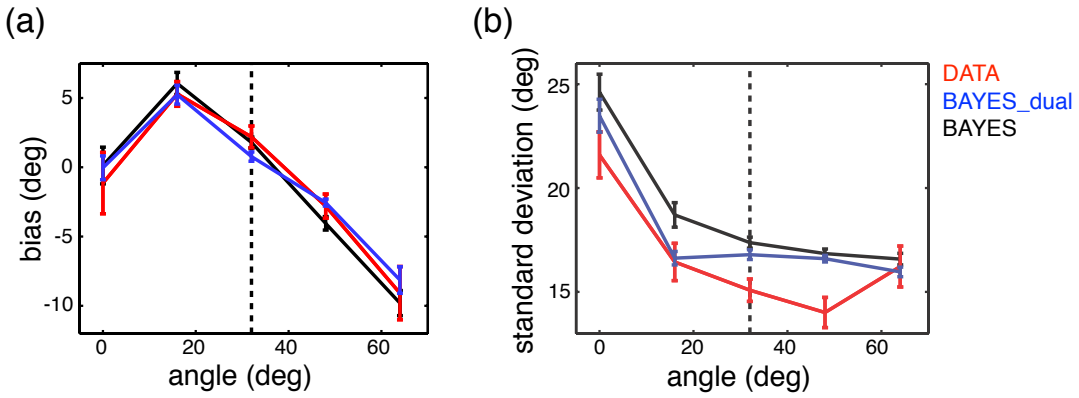


Figure 3.20: Estimation bias (a) and standard deviations (b) obtained with the BAYES_dual model (which also models the detection task; blue) plotted alongside the BAYES model (black), and the experimental data (red). Data points from either side of the central motion direction have been averaged together, so that the furthest left point corresponds to the central motion direction, and the vertical dashed line corresponds to the most frequently presented motion directions (±32°). Results are averaged over all participants and error bars represent the within-subject standard error.

detection performance than was observed experimentally. Overall, these results appear consistent with the hypothesis that participants used a Bayesian strategy to perform the detection task.

The estimation bias and standard deviation predicted by the *BAYES_dual* model is shown in figure 3.20 (blue), plotted alongside the predictions from the *BAYES* model (black) and the experimental data (red). Similar to the *BAYES* model, the *BAYES_dual* model provided a good fit for both participants' estimation biases and standard deviations (mean absolute error of 0.83° & 1.33°, for the fits of the estimation bias and standard deviation respectively; compared with 0.75° & 2.17° obtained with the *BAYES* model).

We considered whether the detection task could have influenced participants' behaviour in the estimation task. For example, the *BAYES_dual* model predicted increased detection performance for the most frequently presented motion directions, so that stimuli perceived as moving in directions that were similar to the frequently presented motion directions would be more likely to be detected. As a result, the size of the attractive estimation bias towards frequently presented motion directions would be increased when we looked only at trials where stimuli were detected.

This interaction between the detection and the estimation task could also have been present in our analysis of the estimation responses of real participants. However, for the *BAYES_dual* model, the detection task was found to have a relatively minor influence on the magnitude

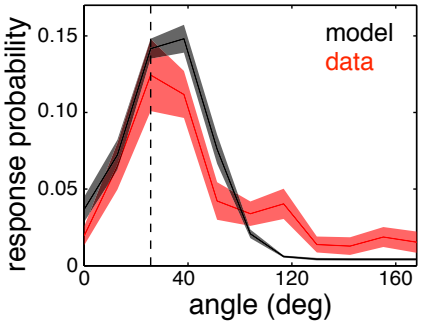


Figure 3.21: Predicted estimation response probability distributions for trials where no stimulus is presented, but where participants reported detecting a stimulus. Model predictions (grey; *BAYES_dual* model, see supplementary materials for details) are plotted alongside the experimental results (red). Data points from either side of the central motion direction have been averaged together in this plot, so that the furthest left data point corresponds to the central motion direction, and the vertical dashed line corresponds to the most frequently presented motion directions (±32°). Results are averaged over all participants and shaded error bars denote the within-subject standard error.

of the estimation bias, for trials where a stimulus was detected (verified by comparing the magnitude of the predicted estimation biases for trials where the stimulus was predicted to be detected, versus the magnitude of the predicted estimation biases for all trials). Therefore, while it is possible that there could have been a small interaction between the two tasks, our modeling work suggests that participants' behaviour in the detection task had a small, and possibly negligible, impact on the experimentally measured estimation biases.

3.3.5.3 Predicted estimation responses in the absence of a stimulus

We were interested to see whether the prior and likelihood distribution that we derived to fit participants' response distributions when a stimulus was presented could explain their estimation behaviour when no stimulus was presented. To do this, we used the *BAYES_dual* model, with parameters fitted to participants' responses when stimuli were present, to predict participants' estimation responses on trials when no stimulus was present.

Figure 3.21 shows the estimation distributions predicted by the *BAYES_dual* model for trials where there was no stimulus present, but where participants detected a stimulus (black), plotted alongside the experimentally measured distribution (red). The average 'zero-stimulus' estimation distribution predicted by the model provided a good fit for the population averaged estimation distributions, with an R^2 value of 0.71. The behaviour of individual participants was also well predicted by the model: the fits for participants' zero stimulus estimation distributions had a positive R^2 value for 8 out of 12 of them. For these participants, the median R^2 value was

0.65 (0.46, 0.83; 25th & 75th percentiles). The fact that the majority of participants' behaviour in the absence of a stimulus could be predicted from their estimation responses in the presence of a stimulus, provides strong support for hypothesis that performed the task using a Bayesian strategy, combining their learned expectations with their sensory evidence using Bayes' rule.

3.4 Discussion

3.4.1 Summary of results

We found that participants quickly and automatically developed expectations for the most frequently presented directions of motion. On trials where no stimulus was presented, but where participants reported seeing a stimulus, they were strongly biased to report motion in the two most frequently presented motion directions (figure 3.5). This bias could not be explained as due to any particular 'response-strategy'. Participants' perception of real motion stimuli was also influenced by their learned expectations: they showed increased detection performance for the most frequently presented motion directions (figure 3.10), and estimated stimuli to be moving in directions that were more similar to the most frequently presented motion directions than they really were (figures 3.7). Participants' estimation behaviour was well described by a model which assumed that they solved the task using a Bayesian strategy, combining a learned prior of the stimulus statistics with their sensory evidence in a probabilistic way (figures 3.12). Finally, our model of participants' behaviour in the presence of a stimulus was able to predict their estimation responses when no stimulus was presented (figure 3.21).

3.4.2 Learning to 'expect' frequently presented motion directions

Participants rapidly learned to expect the likely stimuli within just a few minutes of task-performance. One byproduct of such rapid learning was that because participants learned which motion directions were expected within very few trials, it was difficult for us to measure the short term time-course and dynamics of learning (figure 3.6). In section 5.2 we discuss how our work could be extended to investigate how expectations are acquired over time.

Other psychophysical studies have shown that rapidly learned expectations influence perception of bistable stimuli (Haijiang et al., 2006; Sterzer et al., 2008). In common with our results, these studies found attractive perceptual biases towards participants' expectations. However, while these studies looked at perception of relatively complex visual features, such as whether a stimulus was rotating (Sterzer et al., 2008), our experiment looked at perception of simple unambiguous features, which are likely to be processed at a lower level in the visual hierarchy, such as cortical area MT (Newsome et al., 1989). Whether similar neural changes are responsible for the effects of expectations on perception of both simple and more complicated stimulus features is an open question.

Previous studies have also reported that visual 'hallucinations' can be induced by top-down expectations (Zhaoping and May, 2007; Grossberg, 2000). In particular, our finding, that participants perceived motion in expected directions when nothing was presented (figure 3.5), is similar to what has been reported previously by Seitz et al. (Seitz et al., 2005b). Seitz et al. found that after extended training in a psychophysical task, participants reported seeing dots moving in the trained direction even when no stimulus was displayed. However, an important difference between our results and the experiment of Seitz et al. was the time taken for these hallucinations to develop: in the study of Seitz et al. it took around 8 1hr sessions for participants to perceive motion in the trained direction when there was nothing there, while we observed this effect within the first 200 trials. Thus, given the large differences in timescale, it seems unlikely that the observed no-stimulus biases were produced by the same underlying phenomena in both cases.

3.4.3 Bayesian model

In our experiment, participants learned the statistics of the presented motion directions. In Bayesian terms, this corresponds to learning a prior distribution of the motion stimuli. Bayesian theory tells us how prior knowledge should be combined with sensory inputs to make optimal estimates (Jaynes, 1986) (section 1.2). Our results can thus be interpreted in the context of two questions: 1) are participants able to learn a prior about motion stimuli in the course of our experiment?; 2) is this prior combined optimally with participants' sensory observations to lead to motion estimates?

We constructed a simple model of participants' estimation behaviour, which assumed that on each trial they combined their sensory evidence (based on a noisy sensory measurement of motion direction) with a learned prior distribution of 'expected' motion directions, using Bayes' rule (figure 3.11). This model provided a good fit to participants' estimation biases and standard deviations (figure **??**). Interestingly, the quality of the fit to the data did not decrease when the width of the likelihood was held constant with presented motion direction (figure 3.12). The learned prior (figure 3.18) was found to be qualitatively similar in shape to the true stimulus distribution (figure 3.1), indicating that participants were able to rapidly learn a multi-modal prior over motion direction.

In our experiment, there was a large degree of overlap between the luminance of the two staircased contrast levels (determined by running staircases on the detection performance). Thus, we separated participants' estimation responses into 'low' and 'high' contrast trials, determined by the contrast of each individual trial, rather than the staircased contrast level that they were a part of. We found that the standard deviation in participants' motion direction estimates was largest for low contrast stimuli. In addition, participants exhibited larger estimation biases towards frequently presented motion directions with the low contrast stimuli

(figure 3.2.3.1).

These results are qualitatively consistent with what we would expect if participants behaved as Bayesian observers. At low contrast levels, participants' sensory uncertainty should increase (i.e. corresponding to an increase in the width of their likelihood function). As a result, their learned prior should have a stronger influence on their estimates of motion direction, resulting in increased estimation biases. Unfortunately however, we were not able to well fit participants' estimation behaviour at each contrast level using our Bayesian model, as there there were too few data points per experimental condition to adequately constrain the model. Future work, using more experimental trials for each contrast level, could investigate how the perceptual biases induced by a learned prior vary with stimulus contrast (Stocker and Simoncelli, 2006a).

We reasoned that if participants were indeed behaving as Bayesian observers, then the prior and likelihood derived from their estimation responses when a stimulus was present should predict their estimation behaviour when no stimulus was present. This is indeed what we found: the majority of participants' zero-stimulus estimation distributions were well fitted by the model (figure 3.5). Thus, while 'hallucinating' motion when none is there could potentially be disadvantageous in some everyday situations (Seitz et al., 2005b), in the context of our experiment, it is just what we would expect for an ideal Bayesian observer who sought to minimize their estimation error in the face of perceptual uncertainty.

We compared our Bayesian model with various 'response bias' models, which assume that participants respond according to different strategies on different trials: either relying entirely on their sensory observations, or on their expectations. These models were worse at describing the estimation data than the Bayesian model (larger *BIC* values; figure 3.12), leading us to rule them out as an explanation for participants' behaviour in the estimation task.

Our finding that participants responded according to a 'single-strategy' Bayesian model does not necessarily imply that the biases we observed were perceptual in origin. For example, it is possible that participants altered their overall behavioral strategy in order to incorporate knowledge about which motion directions were most likely, while the perceptual appearance of the presented stimuli remained unchanged. Indeed, distinguishing between biases that occur at the perceptual or decision-making level is a very difficult task to perform psychophysically (Schneider and Komlos, 2008) (see section 2.3.1).

However, our modeling work does suggest that participants' combined their expectations with their sensory observations in a non-trivial way. Specifically, on each trial participants did not rely solely on either their expectations or their sensory observations, but rather, they made their estimations based on a combination of both of these sources of information. Further, we noted that if the observed estimation biases were due to a change in behavioural strategy, this must have occurred at a largely subconscious level, as most participants were unable to indicate the two motion directions that had been most frequently presented, with a large proportion

(9 out of the 12 participants included in our analysis) reporting either that there were equal number of stimuli moving in all directions, or that most of the stimuli were centered around a single motion direction. Also, our personal observations setting up the experiment was that lab personnel often perceived patterns of moving dots in zero contrast trials, leading us to the conclusion that experimental subjects experienced the same "hallucinations".

3.4.4 Eye movements

In the experiment of Sekuler & Ball (Ball and Sekuler, 1982), subjects reported that they experienced their eye movements being involuntarily 'pulled' in the direction of the stimulus. Thus, it was suggested by the authors that mechanisms controlling eye movements might be capable of responding to very low luminance motion stimuli, and thus, that the resulting eye movements could be used by participants to help them correctly detect stimuli that were otherwise imperceptible. Similarly, it is possible that subjects' eye-movements could have contributed to the changes in detection performance and reaction time that we observed. That is, if subjects were biased to move their eyes in 'expected' motion directions, then this could have resulted in decreased detection thresholds for these motion directions.

However, the effect of subjects' eye-movements on their estimates of motion direction is not clear. Naively, we might expect that, if participants were biased to move their eyes in expected motion directions, then this would produce estimation biases *away* from the expected motion directions (as the motion component in this direction would be reduced, relative to the motion of the eye), instead of the attractive estimation biases that we observed. On the other hand, if participants were moving their eyes in the expected motion direction, they could have used proprioceptive or efference copy information to aid their judgements of stimulus motion direction, which would lead to a bias towards the expected motion directions.

To investigate how expectations alter motion perception and eye-movements, Krauzlis et al. conducted an experiment in which subjects were required to report which out of two horizontal directions stimuli were moving in (Krauzlis and Adler, 2001). Interestingly, Krauzlis et al. found that subjects' eye-movements and perceptual decisions tended to be the same on a trial-by-trial basis. Further, they found that a sensory cue indicating the direction that stimuli were most likely to be moving, caused a similar shift in subjects' eye-movements and perceptual decisions (towards the cued direction). On the basis of these results, Krauzlis et al. postulated that prior expectations alter the activity of motion-selective neurons that are read out by both the oculomotor and perceptual systems.

However, it is possible that in the experiment of Krauzlis et al., subjects' eye-movements were biased by the presented cue itself, rather than by their expectations about the stimulus motion direction. This is because the same effect would be observed if there was a a tendency for subjects to look at the presented cue. To control for this effect, it would be interesting

to measure subjects' eye-movements in our experiment, in which subjects' expectations are manipulated implicitly, through the stimulus statistics.

3.4.5 Interaction between tasks

Previously, it has been found that performing a discrimination task on stimulus motion direction can bias participants' estimation responses (Jazayeri and Movshon, 2007). We asked whether a similar interaction between tasks could have occurred in our experiment. Specifically, we asked whether participants estimation biases could have come about as a result of changes in their detection behaviour. To illustrate how this could happen, consider the case where participants' expectations alter their detection responses, without altering their estimation responses. Thus, if participants were more likely to detect a stimulus when they perceived it to be moving in 'expected' directions, their estimation distributions would appear to be biased towards these directions when we looked only at trials where a stimulus was detected. However, this bias would disappear when we looked at estimation responses from all trials, regardless of participants' detection responses, which is not what we find experimentally (there was no significant difference between the estimation biases calculated from trials where participants detected stimuli, and from all trials; $p = 0.71$, 5-way within-subjects ANOVA).

However, if, on trials where participants did not detect a stimulus, they treated the estimation task as meaningless and provided random estimation responses, then on average we would still observe a bias towards the expected directions. This strategy would allow that participants to behave in a 'self-consistent' way in both tasks (Stocker et al., 2006): when they have settled on the hypothesis that there is no stimulus present, it makes little sense for them to scrutinize which direction it is moving in. However, as discussed earlier, participants' detection performance varied relatively weakly with motion direction, with an population averaged difference in detection performance of only 5.9±1.0% between the two most frequently presented motion directions, and other directions (figure 3.10a). Thus, it seems unlikely that the highly significant variation in estimation biases observed experimentally (varying by 14.6±2.9° between stimuli moving at ±16° and ±64°; figure 3.7a) could be brought about by such small changes in detection performance.

3.4.6 Relation to motion-aftereffect illusion

Our finding that subjects exhibited attractive estimation biases *towards* frequently presented motion directions is at odds with the commonly reported 'motion aftereffect' where people exhibit repulsive estimation biases *away* from a motion stimulus presented immediately beforehand (see (Anstis et al., 1998) for a review).

We asked whether the reason that we did not observe any repulsive estimation biases was because the presented stimuli were low contrast, and thus not strong enough to induce sensory

adaptation. To test this, we analyzed subjects' estimation biases on trials directly following a high contrast motion stimulus (section 3.2.3.1). However, we found that even the high contrast stimuli had no effect on subjects' perception of subsequently stimuli: 11 out of the 12 included subjects did not exhibit a significant estimation bias towards or away from the previously presented high contrast stimulus.

Recent analysis of our data indicates that on trials where no stimulus is presented, subjects exhibited a slight tendency to report stimuli moving in the opposite direction from expected, indicating that there may have been a small motion aftereffect (Vincent Valton, personal communication). Note however, that if motion aftereffect illusion caused subjects to occasionally report motion in the opposite direction from the previous stimulus, then this would not qualitatively alter their trial averaged estimation bias on trials where stimuli were presented.

A possible reason that we did not observe a strong motion aftereffect in our data could have been due to the relatively large inter-stimulus-interval ('ISI') which allowed motion selective neurons to 'recover' following short-term adaptation from the previous motion stimulus. This hypothesis is supported by a psychophysical study conducted by Kanai et al. (Kanai and Verstraten, 2005), who found that varying the ISI between an adaptor and test stimulus can produce in qualitatively different perceptual biases: repulsive biases are observed for a short ISI, attractive biases for a long ISI.

In our experiment, subjects' were free to move their eyes during stimulus presentation. Thus subjects' eye-movements could be an alternative or contributing factor to seeing a bias towards the expected direction, rather than a motion aftereffect. Indeed, previous studies have suggested a close relation between subjects' eye movements and the motion aftereffect illusion, with some studies suggesting that the illusion is stronger when eyes are fixated (**???**).

We modelled subjects' behaviour using a simple Bayesian model, which hypothesized that subjects' combined their learned expectation for frequently presented motion directions with their received sensory input using Bayes' law. While our model was able to account for the attractive estimation biases observed in our experiment, it is worth noting that, in its present form our model cannot account for the repulsive estimation biases observed during the motion aftereffect.

In chapter 5 we discuss ways in which Bayesian models could be able to account for repulsive estimation biases observed during motion adaptation. One possibility is that exposure to novel stimulus statistics may cause subjects to alter their *likelihood function* that describes how presented stimuli give rise their observed sensory estimate, as opposed to their *prior*, which describes the probability that different stimuli are presented. Unlike changes in the prior, changes in the likelihood function may lead to repulsive biases away from a frequently presented stimulus (Stocker and Simoncelli, 2006b). In general the timescale over which subjects' alter their prior expectations or their likelihood function will depend on how rapidly they expect different

aspects of their world to change (see section 5.2 for further discussion).

An alternative possibility is that we learn an internal model that describes how local and global image properties combine to generate our received sensory input (Schwartz et al., 2007). Thus, repulsive perceptual biases would emerge because people try to 'factor out' the *global* image statistics (i.e.. the spatiotemporal context) to estimate local image properties (i.e. the *difference* between a stimulus and its spatiotemporal context). We discuss this class of models further in section 5.1.

Chapter 4

Goal-orientated attention as reward-driven optimization of sensory processing

In this chapter, we propose a normative framework for considering why and how attention alters the responses of visual neurons. To account for the effect of behavioural demands on visual processing, we hypothesize that the nervous system learns an internal model that predicts how *both the sensory input and the reward* received for performing different actions are determined by a common set of explanatory causes. We postulate that this internal model is in general imperfect and that attentional processes correspond to its temporary optimization for the task at hand. A simple normative model based on this idea is able to predict a number of task-dependent changes to the responses of visual neurons, including modulation of neural contrast responses functions, sensory tuning curves and center-surround interactions. Our results demonstrate how a diverse range of experimentally observed task-dependent effects are predicted as a result of functional principles, providing a new perspective on previous phenomenological models of goal-directed visual attention.

4.1 Methods

4.1.1 Simulated visual stimuli and behavioural task

In many experimental investigations of goal-directed visual attention, a monkey is instructed (often via a visual cue) that a particular spatial location is 'task-relevant', and thus, should be attended. In order to receive a reward in the task, the animal is then required to make responses that are contingent on stimuli presented at this location, while ignoring distractor stimuli presented at other locations (Luck et al., 1997; Reynolds et al., 2000; Williford and Maunsell,

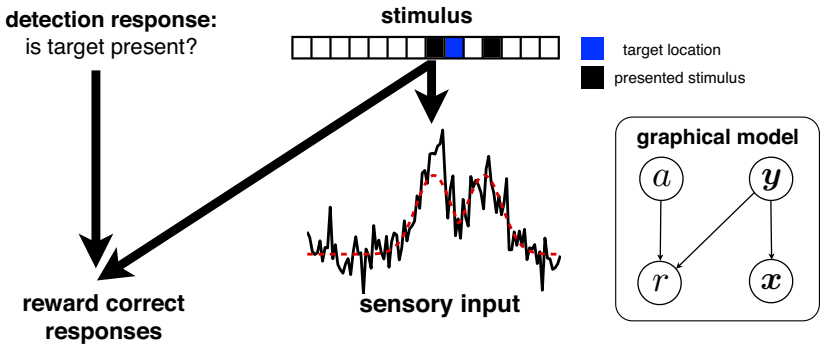


Figure 4.1: Schematic of the detection task. Presented stimuli ($\boldsymbol{y}$) are represented by binary variables, each indicating whether a stimulus is present at a particular location. Stimuli combine to produce the sensory input ($\boldsymbol{x}$; black curve). One or more locations are selected as 'target' locations in the detection task (the target is unknown to the agent at the start of the task). The agent gives a response (a) indicating whether a stimulus is present at a target location, based on their sensory input and their learned model of reward. Correct responses are followed by a reward (r). Inset is the corresponding graphical model of the task.

2006; Roberts et al., 2007). To capture some of the main aspects of these experiments, we simulate a visual detection task, in which an agent is presented with (one or more) stimuli at various locations, and has to report whether a stimulus is present at a single 'target' location (figure 4.1). The agent receives a unitary reward for a correct response in the task, and no reward otherwise. Note that in our simulations the agent is not explicitly told where to attend to: stimuli are equally likely to be presented at all locations, and the agent must learn which locations are behaviourally relevant (i.e. the target location(s)) through feedback in the task.

The sensory input statistics in our model are described by a binary latent variable model (Puertas et al., 2010). Presented stimuli are represented by binary hidden variables ($y_i \in \{0,1\}$), with each variable representing a different spatial location (for example, $y_i = 1$ indicates that a stimulus is present at the i^{th} spatial location). There is an equal probability for stimuli to be presented at all locations, and stimuli are presented at different locations independently of each other:

$$p(\boldsymbol{y}) = \prod_{i=1}^{n_y^{true}} p(y_i), \quad p(y_i = 1) = \alpha. \tag{4.1}$$

where n_y denotes the number of spatial locations, and α denotes the probability that a stimulus is presented at any particular location (α was chosen to be identical to the prediction for $p(y_i = 1)$ given by the agent's internal model, before attentional optimization; see section 4.1.2.1).

Stimuli combine non-linearly to generate the sensory input signal received by the agent ($\boldsymbol{x}$),

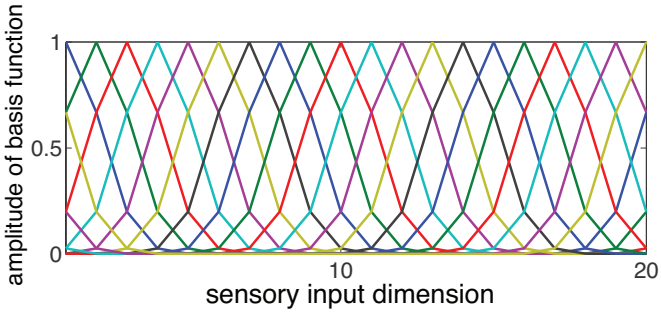


Figure 4.2: Basis functions used to generate the received sensory input (for the initial simulations, where stimuli which included a spatial but not a featural dimension). Each plot shows a single column of the basis function, '$\boldsymbol{A}$'. Individual plots represents the mean sensory input generated by a single active y-unit. Note that the basis used to generate the sensory input are the same as the agent's internal model.

according to:

$$x_i = \max_j \{A_{ij} y_j\} + \gamma_i, \tag{4.2}$$

where γ_i is a Gaussian noise variable (with zero mean and variance σ^2), and $\boldsymbol{A}$ is an $n_x \times n_y$ matrix of basis functions.

The basis functions (columns of $\boldsymbol{A}$) were chosen so that a stimulus presented at a single location activated several neighbouring sensory inputs. Sensory inputs (components of $\boldsymbol{x}$) were labelled with n_x equally spaced values between $-\pi$ and π (producing a vector of spatial locations; '$\tilde{\boldsymbol{x}}$'). Likewise, each of the y-units (components of $\boldsymbol{y}$) was labelled with n_y equally spaced values between $-\pi$ and π (producing a vector of 'preferred' spatial locations; '$\tilde{\boldsymbol{y}}$'). Elements of $\boldsymbol{A}$ were given by:

$$A_{ij} = \exp\left(\frac{-(\tilde{x}_i - \tilde{y}_j + 2\pi k)^2}{2\lambda_A^2}\right). \tag{4.3}$$

where k is an integer, set so that $-\pi < (\tilde{x}_i - \tilde{y}_j + 2\pi k) < \pi$ (so that the stimulus space is circular and there are no edge effects), and λ_A determines the width of the basis functions. Columns of $\boldsymbol{A}$ are plotted in figure 4.2. Each plot can be interpreted as the mean sensory activation produced by a stimulus presented at one particular spatial location.

One or more spatial locations, indexed by I, were chosen as 'target' locations in the task. The detection target ($t \in \{0,1\}$) was classified as present if a stimulus was present at at least one of the target locations ($t = 1$ if $\exists i \in I : y_i = 1$). The agent was required to give a response indicating whether they believed the target stimulus was present or not ($a = 0$ or 1 for a rejection or detection response respectively). They received a unitary reward for a correct response, and no reward otherwise:

$$r = \begin{cases} 0 & \text{if } a \neq t \\ 1 & \text{if } a = t. \end{cases} \tag{4.4}$$

4.1.2 Bayesian model of visual processing and task performance

To perform the task, the agent has to use their received sensory input ($\boldsymbol{x}$) to estimate the reward (r) associated with each possible action (a). We assume that they do this by learning a probabilistic model that describes how both the reward and sensory input are generated by a common set of hidden causes (Sahani, 2004). The assumed goal of the visual system is to compute the posterior distribution over the hidden causes, given the received sensory input. This information, encoded in the firing rates of visual neurons, is then used by the agent to estimate the reward associated with each action. In section 4.1.2.1, we describe the agent's internal model of their sensory inputs; in section 4.1.2.2, we describe their internal model of reward; and in section 4.1.2.3 we simulate the visual neuron responses.

We hypothesize that attentional processes continuously adapt the agent's internal model in order to improve their predictions of the received reward for performing different actions. Section 4.1.2.4 describes in detail how the agent's internal model is optimized towards the task in our simulations.

4.1.2.1 Agent's generative model of sensory inputs

The agent uses a hierarchical internal model to infer the hidden causes of their received sensory input (figure 4.3a). Thus, in contrast to the simulated experiment, where spatially localized stimulus features are presented independently of each other, the agent assumes that there is a higher level of statistical structure, such that certain image features are more likely to be presented together than others (see section 4.1.3 for further discussion).

In the agent's internal model, high-level hidden variables ($\boldsymbol{z}$) are assumed to generate lower-level hidden variables ($\boldsymbol{y}$), which in turn generate the received sensory input ($\boldsymbol{x}$). The joint probability distribution for this model is of the form:

$$p\left(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}|\theta\right) = p\left(\boldsymbol{x}|\boldsymbol{y},\theta\right) p\left(\boldsymbol{y}|\boldsymbol{z},\theta\right) p\left(\boldsymbol{z}|\theta\right), \tag{4.5}$$

where θ denotes the parameters of the agent's internal model of the sensory inputs.

All hidden variables are binary ($y_i \in \{0,1\}$, $z_i \in \{0,1\}$), while the observed data ($\boldsymbol{x}$) are continuous. For mathematical simplicity, we apply the constraint that a maximum of one z-unit can be active at a time, with equal probability associated with all units:

$$p\left(z_i = 1, \boldsymbol{z}_{/i} = \boldsymbol{0}|\theta\right) \propto \rho/n_z, \quad p\left(\boldsymbol{z} = \boldsymbol{0}|\theta\right) = 1-\rho, \tag{4.6}$$

where n_z denotes the number of z-units in the model, and ρ denotes the probability that one of the z-units is on.

Given $\boldsymbol{z}$, the y-units are assumed to be conditionally independent ($p(\boldsymbol{y}|\boldsymbol{z},\theta) = \prod_{i=1}^{n_y} p(y_i|\boldsymbol{z},\theta)$), with a probability of being active given by:

$$p(y_i = 1|\boldsymbol{z},\theta) = \mathrm{sig}\left(\boldsymbol{b}_i^T \boldsymbol{z} - b_{0i}\right), \tag{4.7}$$

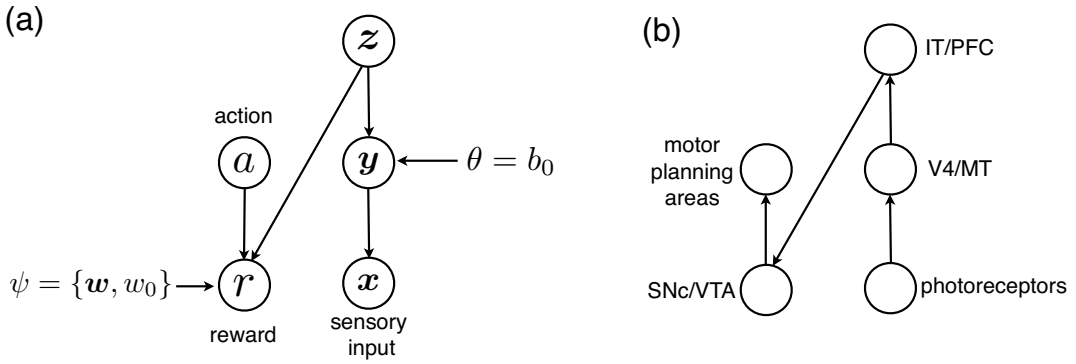


Figure 4.3: Agent's internal model of the sensory input and reward. (a) The agent learns a hierarchical model, where high-level hidden variables (z-units), corresponding to complex spatially distributed image features (e.g. objects/faces), are assumed to determine the state of lower-level hidden variables (y-units), corresponding to simple spatially localized image features (e.g. orientation/motion direction), which generate the received sensory input ($\boldsymbol{x}$). High-level hidden z-variables are also assumed to generate the reward received (r) for performing different actions (a) in the task. During task performance, the agent updates parameters that predict how the reward depends on the high-level hidden variables in their model ($\psi = \{\boldsymbol{w}, w_0\}$), as well as parameters that determine the probability individual y-units are active ($\theta = \boldsymbol{b}_0$) (b) Putative mapping of probabilistic model onto neural architecture. Arrows denote the direction of feedforward processing (both direct and indirect). Incoming sensory signals are first processed in low and intermediate visual areas, such as V4 and MT, before being sent to higher level sensory areas, such as the inferotemporal and prefrontal cortex (IT and PFC). These high-level sensory areas project to regions in the basal ganglia, such as the substantia nigra colliculus (SNc) and ventral tegmental area (VTA), which compute the expected reward for performing different actions.

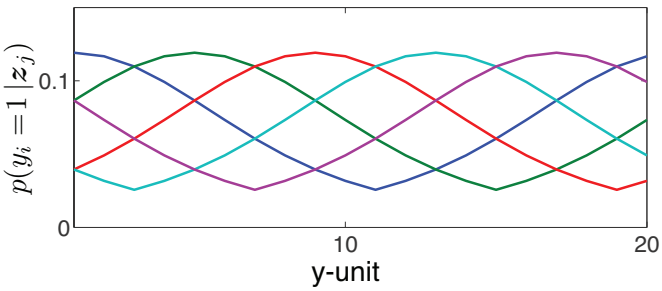


Figure 4.4: Basis functions used for agents' internal model of their sensory inputs in the initial simulations, where stimuli which included a spatial but not a featural dimension. Each plot shows the probability that the agent assumes different y-units are active, given a single active z-unit: $p(y_i = 1|\boldsymbol{z}_j) = \text{sig}(b_{ij} - b_{i0})$ (before task optimization, with bias terms 'b_{i0}' set to their initial values).

where $\text{sig}(x) = (1 + \exp(-x))^{-1}$, $\boldsymbol{b}_i$ is an $n_z \times 1$ basis vector, and b_{0i} is a scalar bias term.

The basis vectors $\boldsymbol{b}_i$ were setup so that when a given z-unit is active, there is an increased probability for neighbouring y-units to be active. Components of $\boldsymbol{z}$ were labelled with n_z equally spaced values between $-\pi$ and π ('$\tilde{\boldsymbol{z}}$'). Elements of $\boldsymbol{b}_i$ were given by:

$$b_{ij} = b_{max} \exp\left(\frac{-(\tilde{y}_i - \tilde{z}_j + 2\pi k)^2}{2\lambda_B^2}\right), \tag{4.8}$$

where λ_B denotes the width of the basis function, and b_{max} determines how strongly z-units determine whether the y-units are on. Figure 4.4 plots the conditional probability that each of the y-units are on, for a given active z-unit.

The agent's internal model that predicts how the sensory input ($\boldsymbol{x}$) is generated by the hidden causes ($\boldsymbol{y}$) was set to be identical to the 'true' data generation process described previously (see equations 4.2 & 4.3).

At the beginning of the task, the true probability that each y-unit was on was set to be equal to the agent's internal model ($\alpha = p(y_i = 1|\theta_{initial})$). Consequently, the only difference between the agent's model and the true model generating the sensory inputs were the second-order statistics describing the probability that different y-units are on at the same time: for the 'true' model, all y-units were independent, while for the agent's internal model there was a higher probability that adjacent y-units were simultaneously active.

4.1.2.2 Agent's generative model of reward

We assume that the agent learns an internal model that predicts how the received reward depends on their performed action, and the state of 'high-level' hidden variables in their internal model (figure 4.3a). Their model of the detection task includes a binary 'target variable'

($t \in \{0,1\}$), that depends on the state of the z-units in their internal model. Given $\boldsymbol{z}$, they assume that the probability of the target being present is:

$$p\left(t=1|\boldsymbol{z},\psi\right)=\mathrm{sig}\left(\boldsymbol{w}^{T}\boldsymbol{z}-w_{0}\right), \tag{4.9}$$

where $\boldsymbol{w}$ (an $n_z \times 1$ vector), and w_0 state how the target variable depends on each of the z-units. A reward of $r=1$ is predicted if $a=t$, and no reward ($r=0$) otherwise (equation 4.4). The agent does not initially know the true location of the detection target: $\boldsymbol{w}$ and w_0 have to be learned online through task-feedback (see section 4.1.2.4).

After receiving a sensory input, the predicted reward for making a detection response is proportional to the posterior probability that the detection target is present (and conversely, for a rejection response, the probability that the target is not present):

$$Q\left(a;\boldsymbol{x},\theta,\psi\right)=\left\langle p\left(t=a|\boldsymbol{z},\psi\right)\right\rangle_{p(\boldsymbol{z}|\boldsymbol{x},\theta)}. \tag{4.10}$$

We assume that the agent makes the response associated with the highest predicted reward. Thus, if the posterior probability that the target is present is greater than 0.5, the agent should make a detection response; otherwise, they should make a rejection response.

4.1.2.3 Visual neuron firing rates

Figure 4.3b illustrates a putative mapping of the probabilistic model used in our simulations onto the neural architecture. The assumed role of the visual system is to infer the posterior probability distribution over the hidden causes. The posterior distribution, encoded in the population activity of visual neurons, is then transmitted to areas of the brain that are responsible for predicting the received reward for performing different actions, allowing the agent to make an appropriate response in the task.

For our simulations, we assume that mean firing rate of a single neuron in the visual cortex encodes the posterior probability that a single hidden cause is active. Thus, the firing rate of the i^{th} visual neuron can be computed directly from Bayes' rule:

$$\begin{aligned} p\left(y_i=1|\boldsymbol{x},\theta\right) &= \frac{p\left(\boldsymbol{x},y_i=1|\theta\right)}{p\left(\boldsymbol{x}|\theta\right)} \\ &= \frac{\sum_{\boldsymbol{y}_{/i}} p\left(\boldsymbol{x}|y_i=1,\boldsymbol{y}_{/i},\theta\right) p\left(y_i=1,\boldsymbol{y}_{/i}|\theta\right)}{\sum_{\boldsymbol{y}} p\left(\boldsymbol{x}|\boldsymbol{y},\theta\right) p\left(\boldsymbol{y}|\theta\right)}, \end{aligned} \tag{4.11}$$

where $\boldsymbol{y}_{/i}$ represents a vector of all the components of $\boldsymbol{y}$, except for the i^{th} component, and the summation is taken over all possible hidden states.

For our simulations, there were sufficiently few latent variables that we were able to perform the summation over the latent states directly. However, if there is a large number of hidden variables, this summation will become intractable, and an approximate algorithm must

be used. Shelton et al. describe a biologically plausible algorithm that could be used to perform approximate inference on a binary latent variable model similar to the one used in our simulations (Shelton et al., 2011; Puertas et al., 2010).

The stimulus selectivity of a given neuron is largely determined by the basis function of the hidden variable that it encodes. In other words, if a neuron encodes for the presence of a hidden variable that is assumed to generate a specific pattern of sensory activation, then this pattern of sensory activation will indicate that the latent variable is active, and the neuron will respond with a high firing rate. The basis functions used in our simulations were spatially localized (figure 4.4), meaning that model neurons were selective for stimuli at a specific spatial location (called their receptive field, 'RF'). The basis functions of the low-level y-units were more spatially localized than the high-level z-units (compare figure 4.4 a & b), the neurons encoding the y-units had smaller RFs than neurons encoding the z-units. Note however, that in general a neuron's RF will not be identical to the basis function of the encoded variable: while basis functions are an invariant property of the generative model, the recorded RF will depend on the type of stimulus that is used to activate the neuron.

4.1.2.4 Attentional optimization

We postulate that the role of goal-orientated visual attention is to alter the agent's internal model so as to improve their predictions of the received reward (at the potential cost of learning a worse internal model of the received sensory inputs). To do this, we adapt the parameters of their internal model (θ and ψ) online, in order to maximize the average log-probability of the received reward. After each trial, model parameters are updated according to:

$$\theta_{new} \leftarrow \theta + \eta_i \partial_\theta l_i(\theta,\psi), \quad \psi_{new} \leftarrow \psi + \eta_i \partial_\psi l_i(\theta,\psi). \tag{4.12}$$

where η denotes the rate of learning. In appendix B, we show that the derivative of the online objective function can be written as follows,

$$\partial_\psi l_i(\theta,\psi) = \langle \partial_\psi \log p(r_i|a_i,s,\psi)\rangle_{p(s|x_i,r_i,a_i,\theta,\psi)} \tag{4.13}$$

$$\partial_\theta l_i(\theta,\psi) = \langle \partial_\theta \log p(s,x_i|\theta)\rangle_{p(s|x_i,r_i,a_i,\theta,\psi)} - \langle \partial_\theta \log p(s,x_i|\theta)\rangle_{p(s|x_i,\theta)}. \tag{4.14}$$

For the parameters to converge on stable values, we used a learning rate that decreased as a function of the trial number, according to: $\eta = \eta_0/(1+i/n_0)$ (where i is the trial number, and η_0 and n_0 are parameters that determine the initial learning rate and how fast it decays, set to 0.05 & 10^4 respectively). Learning was terminated after a fixed number of N trials (set to 10^5), after which the model parameters were seen to converge on stable values.

To simulate the effect of attention on model neuron responses, we compared the responses of model neurons before and after learning. Note that, our focus was on investigating the *effects* of attentional optimization on model neuron responses, rather than the temporal dynamics of

the attentional optimization process itself. Indeed, while we assume that attentional modulation of visual neuron responses are learned online based on feedback in a task, in reality the visual system could be able to switch its attentional state more quickly, based on previous experience in different tasks (see section 4.3 for further discussion).

We postulated that over the short timescales associated with visual attention, only the prior probability that individual hidden y-units are active varies (determined by the bias terms, b_{0i}, in equation 4.7), while other aspects of the internal model are unchanged. The gradient of the objective function used to update b_{0i} is given by:

$$\langle \partial_{b_{0i}} \log p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) \rangle = \langle \mathrm{sig}\left(\boldsymbol{b}_i^T \boldsymbol{z} - b_{0i}\right) - y_i \rangle. \tag{4.15}$$

Note that evaluating this expression only requires computing first-order statistics, such as the mean activation of the y-units. In comparison, updating the basis functions ($\boldsymbol{b}_i$ and $\boldsymbol{A}$) would require computing second order statistics, which are harder to estimate from a limited supply of noisy data.

Because the y-variables are embedded within a hierarchical Bayesian model, changing the bias term, b_{0i} is not directly equivalent to altering the agent's internal 'prior'. However, it will have a similar effect, altering the marginal probability, $p(y_i|\theta) = \int p(y_i|z,\theta)\, p(z_i)\, dz$, without changing the structure of the model (i.e. which y-units are generated by a given active z-unit). Indeed, we note that in a hierarchical model there is not a sharp cut-off between the 'prior' and 'likelihood' function: altering the likelihood that different y-units are activated by a given high-level z-unit (via the basis function, $\boldsymbol{b}_i$) will act to change the agent's 'prior probability' that different y-units are active when z is unknown.

While the agent is assumed to initially know the general structure of the task (i.e. that they receive a unitary reward for detecting a visual target), they do not know in advance where the target is (w_0 and $\boldsymbol{w}$ are both set to zero initially). These parameters ($\psi \equiv \{w_0, \boldsymbol{w}\}$) are learned online on the basis of the reward received for performing different actions. The objective function gradient used to update $\boldsymbol{w}$ and w_0 is given by:

$$\langle \partial_{w_i} \log p(r|a,\boldsymbol{z}) \rangle = \langle z_i \Big(r - \mathrm{sig}\left(\boldsymbol{w}^T \boldsymbol{z} - w_0\right) \Big) \rangle \tag{4.16}$$

$$\langle \partial_{w_0} \log p(r|a,\boldsymbol{z}) \rangle = -\langle r - \mathrm{sig}\left(\boldsymbol{w}^T \boldsymbol{z} - w_0\right) \rangle. \tag{4.17}$$

4.1.3 Summary of model assumptions

In order to make concrete predictions about how attention should alter visual neuron responses, we made had to make certain assumptions about the neural code, agent's internal model, and the modulatory effects of attention. Here we summarize the assumptions that are critical for our results, providing a brief theoretical justification, and outlining how each assumptions influences the results of our simulations.

- **Neural code.** We assume a very simple neural code, in which the firing rate of each visual neuron is directly proportional to the posterior probability that a single latent variable is active. This choice of code is important for our simulations, as it leads to an expression for neural firing rates which has a similar functional form to previous divisive normalization models of neural responses (Reynolds and Heeger, 2009; Carandini et al., 1997). In these models, neural firing rates are evaluated by dividing their feedforward excitatory drive by a suppressive factor that depends on the summed activity nearby neurons. In our model, divisive normalization emerges from Bayes' law (equation 4.11), due to the fact that the posterior probability for a hidden variable to be active ($p(y_i = 1|\boldsymbol{x})$) is evaluated by dividing the joint distribution over hidden and observed variables ($p(y_i = 1, \boldsymbol{x})$) by the marginal probability for the observed sensory input ($p(\boldsymbol{x})$). In section 5.3 we compare and contrast the neural code used in our simulations to previously proposed neural codes.

- **Sparse stimulus statistics.** We assume that the agent learns a 'sparse' internal model, in which there is a small prior probability for any particular hidden cause to be active (i.e. $p(y_i = 1|\theta) \ll 1$). The theoretical justification for this comes from natural image statistics, which are well accounted for by sparse models (Berkes et al., 2007; Olshausen and Field, 1996). In our simulations, the sparse prior produces strong competition between different explanations of the received sensory input, which is reflected in the divisive suppression of neural responses (see section 4.2.4).

- **Non-linear stimulus combination rule.** The agent learns an internal model in which stimuli are assumed to combine non-linearly, according to a 'max' combination rule (equation 4.2). Indeed, while many previous generative models of visual processing have assumed a linear combination rule, arguably, a strongly nonlinear 'max' combination rule provides a better description of how features combine in natural images (Lücke and Sahani, 2008). In our simulations, a nonlinear combination rule was required to produce neural responses that saturated below their maximum values when the stimulus contrast was high (see figure 4.7). To see why this is the case consider what happens when the agent receives a high amplitude sensory input (i.e. components of x have high values). A high amplitude sensory input can be well explained by a linear model (where $x_i = \sum_j A_{ij} y_j + \gamma_i$) if multiple hidden units are active. Thus, if the agent uses a linear internal model, they will ascribe a high posterior probability for multiple hidden to be active simultaneously (i.e. the posterior probability, $p(y_i = 1|\boldsymbol{x})$, will saturates at ~ 1). However, with a nonlinear 'max' rule ($x_i = \max_j \{A^{true}_{ij} y_j\} + \gamma_i$), multiple hidden variables do not combine to produce a higher amplitude sensory input. Thus, different hidden variables compete to explain the data. Thus, if the agent uses an internal model with a 'max'

combination rule, the posterior probability that they ascribe to individual hidden units will remain below 1, even when the amplitude of the sensory input is high.

- **Internal model structure.** The agent learns a hierarchical internal model, in which high-level hidden variables ($\boldsymbol{z}$), that correspond to the global structure of the sensory input, are assumed to determine the state of low-level hidden ($\boldsymbol{y}$), that correspond to the local features of the sensory input. This model structure is designed to reflect the structure of natural images, in which complex objects are made up of simple image features, which give rise to the observed sensory input (Karklin and Lewicki, 2005; Reichert et al., 2011a). The agent assumes that only the high-level latent variables determine the reward that will be received for performing different actions. This model structure could allow the agent to quickly adapt to new behavioural contexts, as the action that they should perform in any given task will depend on a relatively few number of high-level hidden variables. However, in our simulations, we investigate a situation when this internal model structure is suboptimal: when the image features that are relevant to the task are more spatially localized than the high-level features in the agent's internal model. In our work, it is this mismatch between the structure of the agent's internal model and the behavioural task that drives attentional modulation of visual neuron responses.

- **Attention only alters 'bias-terms' in the internal model.** We postulate that over the short timescales associated with visual attention, the prior probability that individual hidden variables are active can vary (determined by the bias terms, $\boldsymbol{b}_0$; although see previous section), while the image features represented by the latent variables (determined by the basis functions) is fixed. As a result, attention modulates the responses of model neurons to presented stimuli, but does not fundamentally change their stimulus selectivity. Our assumption can be justified functionally from the fact that updating the bias terms only requires estimating first-order statistics, which can be evaluated quicker and more reliably than the second-order statistics that are required to update the basis functions (see 4.1.2.4). More generally, how much attention should alter different aspects of the agent's internal model will depend on several different factors, including the rate at which different aspects of the world vary (discussed in section 5.2.3), and the trade-off between short-term optimization in a specific task, and generalization across many different tasks (discussed in section 5.1.2).

- **Binary latent variable model.** We modeled the stimulus statistics using binary latent variable model. We chose this form of model for simplicity, in order to simulate a simple task where where subjects were rewarded for correctly detecting a stimulus. Recent work suggests that a binary latent variable model also provides a good description of natural scene statistics (Puertas et al., 2010). However, it is worth considering how this

choice of model could have affected our simulations of neural responses as a function of varying stimulus contrast. In figure 4.7 we plot the responses of model neurons while continuously varying the amplitude of the sensory input. As stimulus contrast is not represented by the agent, increasing the amplitude of the sensory input is interpreted as increased evidence that a binary latent variable is 'on', and the model neuron response (given by $r_i \propto p(y_i = 1|\boldsymbol{x})$) increases towards a saturating value. In a more sophisticated model, the agent could learn a joint distribution describing both the probability that a stimulus is present and its contrast. Now, if we assume that the primary goal of the early visual system is to encode the components that make up an image, we should integrate over all possible stimulus contrasts to recover the distribution $p(y_i = 1|\boldsymbol{x})$. Intuitively, such a model should produce qualitatively similar results - a high amplitude sensory input will still indicate a high probability that a stimulus is present. However, future theoretical work will be required to see whether this is true, and thus whether our results generalize to an internal model that includes contrast as a latent variable.

4.2 Results

4.2.1 Attentional modulation of neural population response

The responses of visual neurons to a given visual stimulus can be manipulated experimentally via the presented stimulus statistics (determining which stimuli are *expected*; often communicated via visual cues) (Posner et al., 1980), or the reward delivered for performing different actions in a task (determining which stimuli are deemed relevant to the task, and thus *attended* to) (Pestilli and Carrasco, 2005). Previous 'task-independent' Bayesian models of visual processing, in which the internal model is learned and adapted based on the stimulus statistics alone, can only deal with the first case (Hyvärinen et al., 2005; Simoncelli and Olshausen, 2001). In contrast, the framework presented here, where the agent learns an internal model of both the stimulus and the reward statistics, can account for both stimulus and reward-dependent changes to visual processing. Here, we focus on the latter case, with stimuli equally likely to be presented at all locations, but where only stimuli at certain locations are relevant in determining the actions that the agent must perform to receive a reward.

For the agent's internal model of the sensory input statistics to be altered by the reward structure of the task, there must be some mismatch between their internal model and the external environment (i.e. if the internal model is already a perfect description of the world, it cannot be further optimized). We postulate that, due to the complexity of real-world environments, this will often be the case. In our simulations, we assume that the image features that are relevant to the task (which correspond to the y-units in the agent's internal model) are more spatially localized than the image features used by the agent to choose which action to perform (the

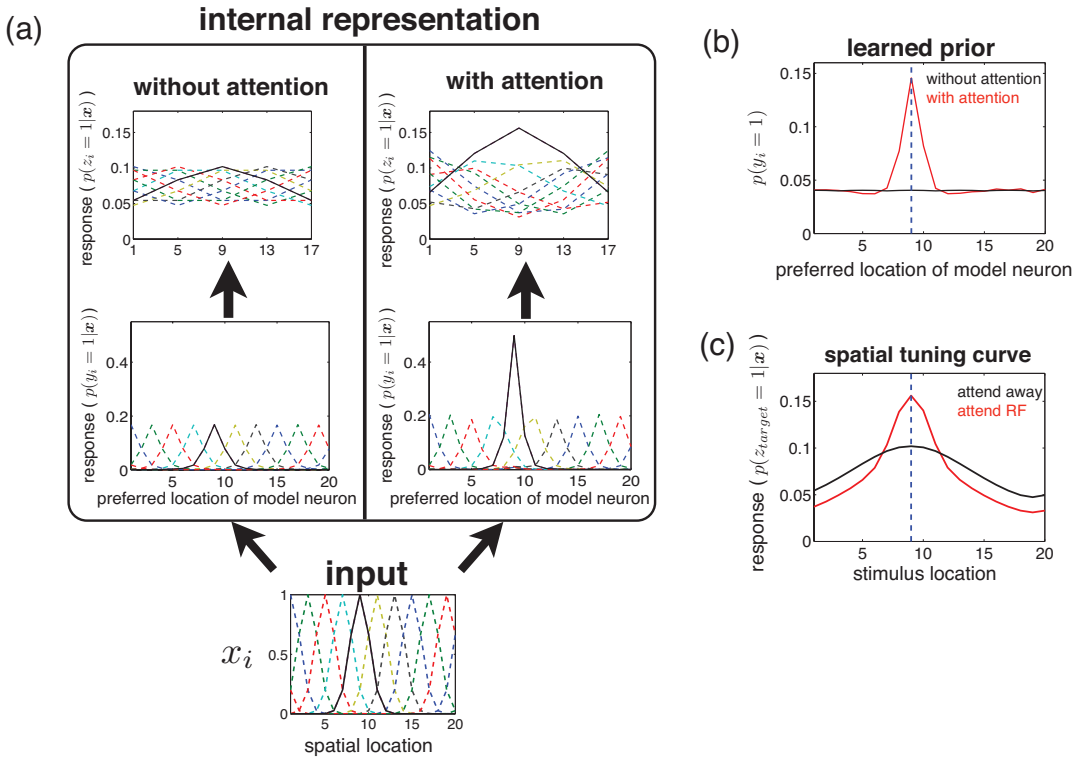


Figure 4.5: Influence of spatial attention on neural population response. (a) Bottom panel: each plot corresponds to the sensory input generated by a stimulus presented at a single location. Sensory input produced by a stimulus at the target location is plotted with a solid black line. For clarity, sensory inputs are generated without noise. Box: each plot corresponds to the neural population response at low (bottom) and high (top) levels of visual processing with a stimulus presented at a single location, without (left) and with (right) attention directed towards a central target location. (b) Prior probability that each of the low-level hidden causes are active, without (black) and with (red) spatial attention directed towards the target location (vertical dashed line). (c) Response of a high-level neuron, plotted as a function of the presented stimulus location, without (black) and with (red) attention directed towards its preferred location (dashed line).

z-units in their internal model). Such a model mismatch might occur because the agent tries to learn a simple model of the behavioural task, in which the action's that they should perform depend on a limited number of spatially distributed image features. While useful in allowing the agent to quickly learn new tasks, this model structure could result in suboptimal performance in experiments that use very simple and/or spatially localized stimuli (e.g. orientated gratings, or coherent motion).

We now describe how the responses of visual neurons in our model towards a stimulus presented at a single location are modulated as a result of attentional optimization towards a detection task. Note that the image statistics and behavioural task used in our simulations are highly simplified, and are designed to provide a 'proof-of-principle' as to how task-dependent modulation of neural responses can be modelled within a Bayesian framework.

As discussed previously (section 4.1.2.3), when a stimulus is presented at a specific spatial location (figure 4.5a, bottom panel, below box), it activates a small number of 'mid-level' visual neurons (corresponding to y-units in the agent's internal model) that are selective for stimuli at this location (figure 4.5a, bottom left panel inside box). In contrast, 'high-level' visual neurons (corresponding to z-units in the agent's internal model), with larger receptive fields (RFs) respond similarly to stimuli presented at many different locations (figure4.5a, top left panel inside box).

The agent uses the responses of high-level visual neurons to choose which action to perform. Because the activity of these neurons does not vary strongly with the presented stimulus location (figure 4.5c, black), the agent will perform suboptimally at detecting whether a stimulus is present at a particular task-relevant location (see next section).

To simulate the effects of goal-directed attention, internal model parameters were learned online to optimize the agent's performance in the task (see section 4.1.2.4). After optimization, the agent learned to associate an increased prior probability for a stimulus at the target location (figure 4.5b). As a result, the sensitivity of mid-level neurons tuned towards this location was increased (figure 4.5a, bottom right panel inside box), and high-level neurons became more selective for stimuli presented at the target location (figure 4.5a, top right panel inside box).

While the agent's learned prior did not match the true stimulus statistics – in reality, stimuli were equally likely to be presented at all locations – after attentional optimization the firing rate of high-level visual neurons was a better predictor of whether a stimulus was present at the task-relevant spatial location, allowing the agent to improve their performance in the task.

4.2.2 Behavioural performance

In our model, the agent chooses which action to perform based on the inferred posterior probability that the detection target is present, given their sensory input ($p(t = 1|\boldsymbol{x})$). We wanted to quantify how well they are able to discriminate whether the target was present, independently

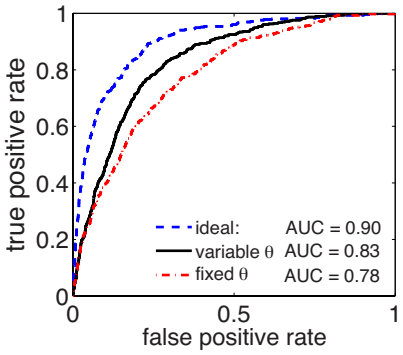


Figure 4.6: Receiver operating characteristic curves, indicating how well the agent is able to classify whether the detection target is present or not. There are three conditions: the ideal case (true model underlying task; blue); a fixed model of how the hidden causes generate the sensory input, but a variable model of how they generate the detection target (fixed θ, variable ψ; red); a variable model of how the hidden causes generate both the sensory input and the detection target (variable θ, variable ψ; black). AUC values (area under the ROC curves) give a summary statistic of how well the agent is able to classify whether the detection target is present or not.

of the reward associated with correct detections or rejections. To do this, we plotted receiver operating characteristic curves (ROC curves; figure 4.6) directly from the agent's estimates of $p(t|\boldsymbol{x})$ (obtained from 2×10^4 simulated trials). The area under the ROC curve (AUC) provides a measure of performance that is independent of the threshold used for classification (Fawcett, 2006).

We consider three conditions: (i) the ideal observer, using the 'true' model underlying the task (blue dashed curve); (ii) where the agent can alter their internal model of how the hidden causes generate the detection target (i.e. ψ can vary), but not how they generate the sensory input (i.e. θ is fixed; figure 4.6, red dashed curve) and (iii) where the agent can alter their internal model of how the hidden causes generate both the reward and the sensory input (variable ψ and θ; figure 4.6, black dashed curve).

The worst performance is obtained when the agent cannot alter θ ($AUC_{fixed} = 0.78$). Varying θ allows the agent to improve their performance above this worst case ($AUC_{variable} = 0.83$), although their performance is still worse than ideal ($AUC_{ideal} = 0.90$). It is unsurprising that the agent cannot perform optimally in our simulations, as we imposed a strong mismatch between their internal model and the true model underlying the task (compare figure 4.3a and figure 4.1, inset). As we assume that they are only able to alter their internal model by changing the prior probability that individual hidden units are on (via $\boldsymbol{b}_0$), they are not able to improve their model so that it is the same as the true model underlying the task. However, they are able improve

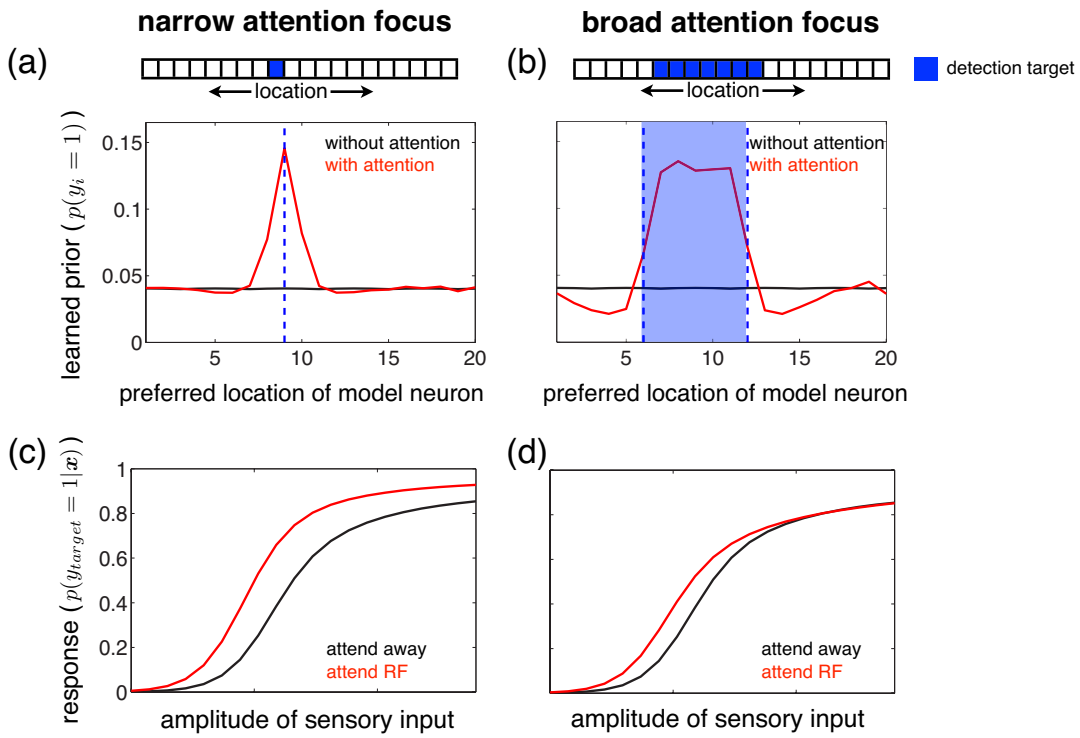


Figure 4.7: Attentional modulation of neural contrast response function. (top panels) Prior probability assumed by the agent that each of the hidden causes are active, without attention (black), or with either a narrow (left) or a broad (right) focus of attention. The attended spatial region is represented by the blue shaded area. (bottom panels) Model neuron response, as a function of the amplitude of a sensory input at the preferred location, without attention (black) or with either a narrow (c) or a broad (d) focus of attention (red).

their performance so that it is better than the worst case scenario, with fixed θ.

How strongly goal-directed attention improves behavioural performance will depend on many factors which determine how well the internal model in the brain can account for the behavioural task. In the work presented here, our focus is to understand how visual neuron response properties are altered by the behavioural context of presented stimuli, rather than how attention alters task-performance.

4.2.3 Attentional modulation of neural contrast response function

There have been a number of controversies about how goal-directed attention alters sensory neural responses. A prominent example is attention-dependent changes to the firing rates of V4 neurons with varying stimulus contrast. Previous experiments have reported very different findings: Williford et al. observed a 'response gain' effect, with increases in neural firing rates for all stimulus contrasts (Williford and Maunsell, 2006), while Reynolds et al. observed a

'contrast gain' effect, consistent with an increase in the effective stimulus contrast (Reynolds et al., 2000) (figure 2.4). Reynolds & Heeger proposed a phenomenological model to account for these differences, arguing that they were due to variations in the relative size of the focus of attention and the stimulus (Reynolds and Heeger, 2009): a narrow focus of attention would give rise to a response gain effect, while a broad focus of attention would give rise to a contrast gain effect (see section 2.3.2). We use our normative model to ask *why* attention might alter neural responses in this way.

To manipulate the size of the attentional focus, we varied the number of target locations in the detection task. We simulated two experimental conditions: one with a single target location ('narrow attentional focus'), and another, with multiple neighbouring target locations ('broad attentional focus'). When the agent optimized their internal model towards a detection task with one target location, they learned to associate an increased prior probability that hidden causes representing stimuli at this location were active (figure 4.7a). With multiple targets, there was a broader change in their learned prior, with increases in the prior probability for hidden causes representing all of the target locations (figure 4.7b).

In our simulations, V4 neurons correspond to hidden variables at an intermediate level of the agent's internal model (i.e. components of $\boldsymbol{y}$). To obtain neural contrast response functions (CRFs), we plotted the mean firing rate of a model neuron while varying the amplitude of a sensory input centred at its preferred location ($\boldsymbol{x} = c \times \boldsymbol{a}_i$, where $\boldsymbol{a}_i$ is the i^{th} column of $\boldsymbol{A}$, and 'c' represents the stimulus contrast). The resulting CRF was qualitatively similar to experiment, increasing monotonically at intermediate sensory input amplitudes, before saturating at high amplitudes. The effect of spatial attention was also consistent with experiment: directing a narrow focus of attention towards the presented stimulus location increased the response of a neuron tuned to this location for all sensory input amplitudes; a broad focus of attention only increased the response of this neuron at intermediate sensory input amplitudes (figure 4.7c & d respectively).

Why does attention alter CRFs as it does in our model? In our work, neural firing rates have a direct functional meaning; they represent the probability that different hidden causes are responsible for producing the observed sensory input. Therefore, we can understand the effects of attention on neural responses directly in terms of how it alters the probability that the agent accords to different 'explanations' of the sensory input (figure 4.8).

When the amplitude of the presented sensory input is low, directing either a narrow or a broad focus of attention towards the presented stimulus location increases the inferred probability that a hidden cause representing this location is active, while decreasing the probability that no hidden causes are active (figure 4.8a & b). Consequently, the firing rate of a model neuron tuned to a low amplitude sensory input is increased by both a narrow and a broad focus of attention (figure 4.7c & d).

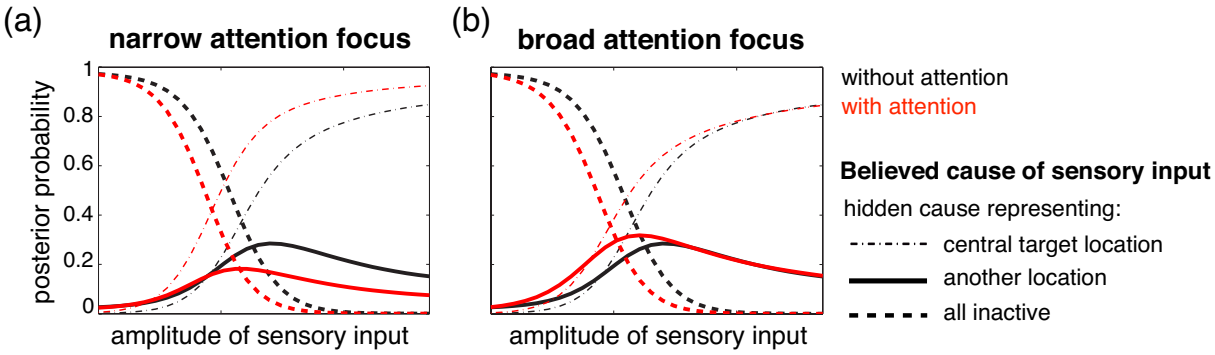


Figure 4.8: Possible explanations of the sensory input, as a function of its amplitude. We consider three types of explanation: those where (i) the hidden cause representing the target location is active; (ii) another hidden cause is active; (iii) all hidden causes are inactive. The probability accorded to each explanation is plotted against the amplitude of a sensory input centred at the target location, without attention (black) or with a narrow (a) or a broad (b) focus of attention directed towards this location (red). At small amplitudes, both sizes of attentional focus increase the inferred probability that the hidden cause representing the attended location is active, reflected by an increase in the response of a model neuron tuned to this location. At large amplitudes, varying the size of the focus of attention produces qualitatively different effects. A narrow attentional focus (a) increases the inferred probability that a hidden cause representing the target location is active, while decreasing the probability that other hidden causes are active. A broad attentional focus (b) leaves the probability associated with these two competing explanations unchanged. Thus, at high sensory input amplitudes, the response of a model neuron tuned to the attended location is increased for a narrow, but not for a broad focus of attention.

In contrast, when the amplitude of the sensory input is high, there is negligible probability accorded to the possibility that no hidden causes are active. In this case, a narrow attentional focus increases the inferred probability that a hidden cause representing the presented stimulus location is active, while decreasing the probability that other hidden causes are active (figure 4.8a). A broad attentional focus, on the other hand, does not alter the probability associated with these two competing explanations (figure 4.8b). Consequently, the firing rate of a model neuron tuned to a high amplitude sensory input is increased by a narrow, but not by a broad focus of attention (figure 4.7c & d).

4.2.4 Relation to 'normalization model of attention'

The predictions of our model are qualitatively similar to the 'normalization model of attention', proposed by Reynolds & Heeger (Reynolds and Heeger, 2009) (described in section 2.3.2). Understanding how their model relates to ours requires writing out the expression for neural firing rates, which are evaluated using Bayes' rule:

$$p\left(y_i = 1|\boldsymbol{x}\right) = \frac{p\left(\boldsymbol{x}, y_i = 1\right)}{p\left(\boldsymbol{x}\right)} \tag{4.18}$$

$$= \frac{\sum_{\boldsymbol{y}_{/i}} p\left(\boldsymbol{x}|y_i = 1, \boldsymbol{y}_{/i}\right) p\left(y_i = 1, \boldsymbol{y}_{/i}\right)}{\sum_{\boldsymbol{y}} p\left(\boldsymbol{x}|\boldsymbol{y}\right) p\left(\boldsymbol{y}\right)}, \tag{4.19}$$

where $\boldsymbol{y}_{/i}$ represents a vector of all the components of $\boldsymbol{y}$, except for the i^{th} component.

We assume that the agent learns a sparse model, with a small prior probability that any particular y-unit is active. Therefore, we can approximate the previous expression by discounting all hidden states where more than one y-unit is active at the same time:

$$p\left(y_i = 1|\boldsymbol{x}\right) \sim \frac{p\left(\boldsymbol{x}|\boldsymbol{y}_i\right) p\left(\boldsymbol{y}_i\right)}{p\left(\boldsymbol{x}|\boldsymbol{y}_0\right) p\left(\boldsymbol{y}_0\right) + \sum_{j=1}^{n_y} p\left(\boldsymbol{x}|\boldsymbol{y}_j\right) p\left(\boldsymbol{y}_j\right)}, \tag{4.20}$$

where $\boldsymbol{y}_i$ denotes a hidden state with only one active y-unit (i.e. $\boldsymbol{y}_i \equiv (0,\ldots 0,1,0\ldots 0)$ with only $y_i = 1$), and $\boldsymbol{y}_0$ denotes a hidden state with all y-units inactive (i.e. $\boldsymbol{y}_0 = \boldsymbol{0}$). We can rewrite this expression in the form:

$$p\left(y_i = 1|\boldsymbol{x}\right) \sim \frac{\mathcal{A}_i E_i\left(\boldsymbol{x}\right)}{1 + \sum_{j=1}^{n_y} \mathcal{A}_j E_j\left(\boldsymbol{x}\right)}, \tag{4.21}$$

where $E_i\left(\boldsymbol{x}\right) \propto \exp\left(\frac{\boldsymbol{a}_i^T \boldsymbol{x}}{\sigma^2}\right)$ and $\mathcal{A}_i = \frac{p(\boldsymbol{y}_i)}{p(\boldsymbol{y}_0)}$. Attention alters the prior probability that the individual y-units are on, increasing the value of $\mathcal{A}_i$ for neurons that are tuned to attended stimuli. $E_i\left(\boldsymbol{x}\right)$ is determined by the sensory input alone, and does not depend on the attentional state of the agent.

The numerator in equation 4.21 ('$\mathcal{A}_i\left(\boldsymbol{b}_0\right) E_i\left(\boldsymbol{x}\right)$') represents the 'excitatory component' of activity, and depends on the dot product between the model neuron basis function and the presented sensory input ($\boldsymbol{a}_i^T \boldsymbol{x}$), as well as the prior probability that the i^{th} hidden unit is active

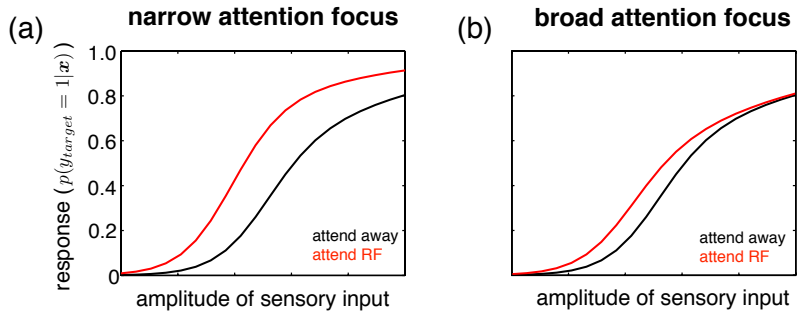


Figure 4.9: Model neuron contrast response functions, approximated using equation 4.21, which only considers hidden states with a maximum of one active y-unit. The model neuron response is plotted as a function of the amplitude of a sensory input at its preferred location, without attention (black) or with either a narrow (c) or a broad (d) focus of attention (red) directed towards this location.

(closely related to $\mathcal{A}_i(\boldsymbol{b}_0)$). The response of the model neuron is suppressed by a divisive normalization factor, '1+1 $+\sum_{j=1}^{n_y}\mathcal{A}_j E_j(\boldsymbol{x})$', which depends on the summed excitatory activity over the population of model neurons. In our simulations, attending to the preferred location of the i^{th} model neuron alters $\boldsymbol{b}_0$, so as to increase the prior probability that the i^{th} hidden unit is active. The resulting change to $\mathcal{A}_i(\boldsymbol{b}_0)$ produces a multiplicative scaling of the excitatory term in the numerator of equation 4.21, as well a change in the degree of suppression from other model neurons (through changes to the divisive normalization factor).

Equation 4.21 is closely related to the expression for the neural firing rates proposed by Reynolds & Heeger in their 'normalization model of attention' (Reynolds and Heeger, 2009). In their model, the response of a neuron with RF centred at a location x is given by:

$$f(x) = \frac{\mathcal{A}(x)E(x)}{\sigma + S(x)}, \tag{4.22}$$

where $E(x)$ represents the excitatory component of the response, $\mathcal{A}(x)$ represents the attention field, $S(x)$ represents the suppressive component of the response, obtained by summing the excitatory activity of neurons across many locations ($S(x) = s(x) * (\mathcal{A}(x)E(x))$, where '$*$' denotes a convolution), and σ is a constant, determining the slope of the contrast response.

To see how well the responses of the model neuron responses are approximated by equation 4.21 we used it to evaluate the model neuron contrast response functions, with either a narrow or a broad focus of attention directed towards the presented stimulus location (figure 4.9). The CRFs obtained under this approximation are very similar to the CRFs obtained using the full posterior distribution (figure 4.7). In particular, the effect of attention is qualitatively similar in both cases, with a narrow focus of attention increasing the model neuron response at all stimulus contrasts, while a broad focus of attention only increases the model neuron response

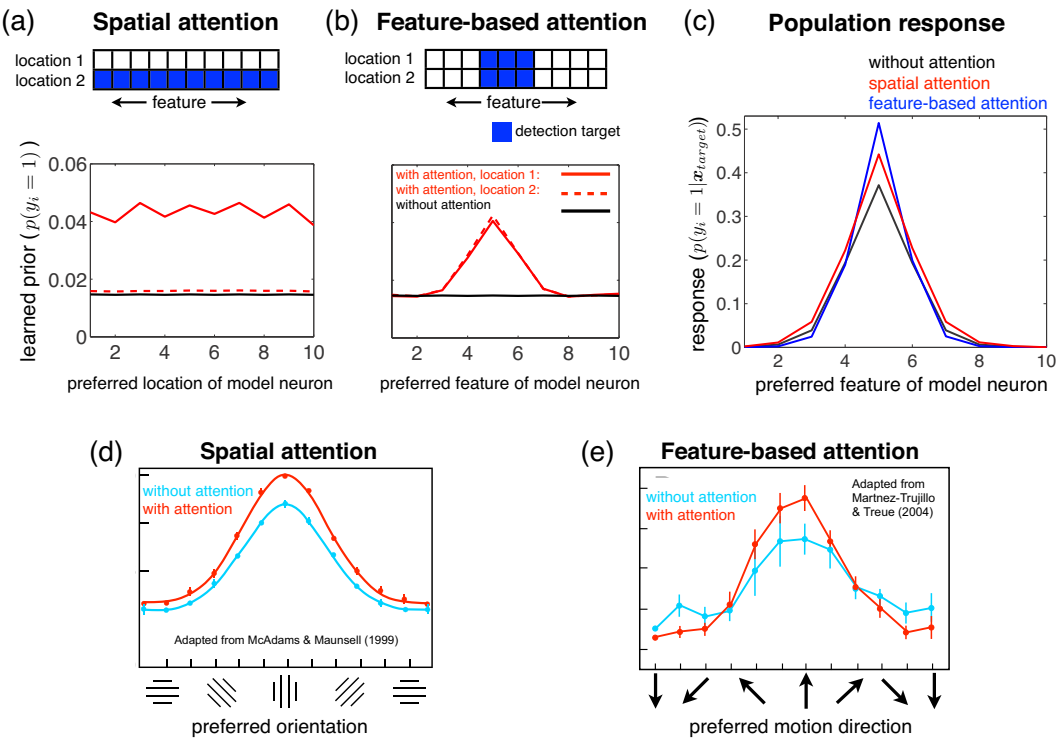


Figure 4.10: Influence of spatial and feature-based attention on the population response. (a & b) Prior probability assumed by the agent that each of the hidden causes are active, without attention (black), or with spatial (a) or feature-based (b) attention. (c) Neural population response in the absence of attention (black), or with attention directed towards the presented stimulus feature (blue) or spatial location (red). (d) Average neural firing rate of a population of V4 neurons, with (red) and without (blue) spatial attention directed towards the RF (adapted from McAdams & Maunsell, 1999). (e) Average neural firing rate of a population of MT neurons with (red) and without (blue) feature-based attention directed towards the presented motion direction (adapted from Martinez-Trujillo and Treue, 2004).

neuron at intermediate contrasts.

4.2.5 Attentional modulation of sensory tuning curves

We investigated how attention alters neural tuning curves in our model. To do this, we extended our model to include both a featural and a spatial dimension. We altered the basis functions that determined the image features represented by the hidden units, so that each model neuron (corresponding to a component of $\boldsymbol{y}$) was selective to both a stimulus feature (e.g. orientation, or motion direction) and a spatial location.

Every sensory input (component of $\boldsymbol{x}$) was allocated a 'feature' label (consisting of 2 lists of $n_x/2$ equally spaced values between $-\pi$ and π; $\tilde{\boldsymbol{x}}_1$), and a 'spatial' label ($n_x/2$ lists of 2 spatial locations, $(0,\pi)$; $\tilde{\boldsymbol{x}}_2$). The y-units were labelled in the same way: each with a corresponding

spatial location and feature ($\tilde{\boldsymbol{y}}_1$ and $\tilde{\boldsymbol{y}}_2$ respectively). Elements of $\boldsymbol{A}$ were given by:

$$A_{ij} = \exp\left(\frac{-(\tilde{x}_{i1} - \tilde{y}_{j1} + 2\pi k)^2}{2\lambda_{ftr}^2} + \frac{-(\tilde{x}_{i2} - \tilde{y}_{j2} + 2\pi k)^2}{2\lambda_{spt}^2}\right), \tag{4.23}$$

where λ_{ftr} and λ_{spt} are parameters determining the width of the basis function along feature and spatial dimensions respectively. The basis functions for the z-units were calculated in the same way as for the previous simulations (see section 4.1.2.1), with all z-units allocated a feature but not a spatial label (i.e. $\tilde{y}_i$ was replaced by the 'feature' label, $\tilde{y}_{i1}$). We set, $\lambda_{ftr} = 1.2$ & $\lambda_{spt} = 2$ (see next section for discussion of how we set λ_{spt}). We also increased the model sparsity, setting $\rho = 0.3$ & $b_{max} = 2$ (so that $p(y_i = 1|\theta_{init}) \approx 0.02$). This increase in sparsity was required to produce robust surround suppression for the simulations described in the next section (but was not critical for simulating the feature tuning curves).

We simulated two experimental conditions. In the first condition ('spatial attention'), one of two spatial locations was selected as a target in the detection task. In the second condition ('feature-based attention'), only certain features were chosen as targets. Spatial attention caused the agent to associate a high prior probability that hidden variable representing the attended location were active, but a uniform prior probability that hidden variables representing different features were active (figure 4.10a). Conversely, feature-based attention caused the agent to associate a high prior probability that hidden variables representing attended features were active, but a uniform prior probability that hidden variables representing both spatial locations were active (figure 4.10b).

Attending towards the presented stimulus location increased the responses of neurons tuned to this location, with no sharpening in the population response (figure 4.10c, red). Similar effects have been observed experimentally in visual area V4 when attention is directed towards a particular spatial location (McAdams and Maunsell, 1999). In contrast, we found that attending to the presented stimulus feature produced a sharpening in the population response; the responses of model neurons that were selective for the attended feature showed increased most strongly by attention (figure 6c, blue). Martinez-Trujillo & Treue reported a similar effect in visual area MT when animals were directed feature-based attention towards a particular motion direction (Martinez-Trujillo and Treue, 2004).

Also consistent with the experimental findings of Martinez-Trujillo & Treue, our model predicted a small suppression in the responses of model neurons tuned to 'unattended' features (Martinez-Trujillo and Treue, 2004). In our model, this suppression came about because the agent accorded greater probability to the possibility that the sensory input was produced by hidden causes representing attended features, at the expense of a reduction in the probability that it was produced by hidden causes representing other, unattended, features.

A notable difference between our simulation results and the experimental data shown in figure 4.10c & d is that the firing rate of neurons that are unselective for the presented stimulus

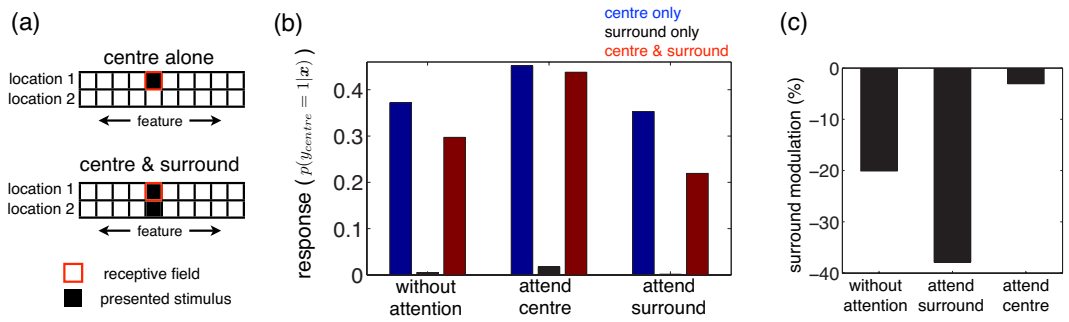


Figure 4.11: Attentional modulation of centre-surround suppression. (a) Schematic of test stimuli. Neural responses were measured with either a single stimulus presented at their RF (top) or with stimuli presented at both their RF centre and surround (bottom). (b) Response of a model neuron to a stimulus presented in the RF centre (blue), surround (black), or at both the RF centre and surround (red), without attention, or with attention directed towards the RF centre (attend centre) or surround (attend surround). (c) Fractional change in model neuron response when a second stimuli is presented in the surround, for each of the three attentional conditions.

feature is zeros, while experimentally, they exhibit at a non-zero baseline firing rate. Our model could be altered to produce this behaviour by altering the structure of the agent's internal model so that the mean sensory activation is non-zero when no stimulus features are present (i.e. when $\boldsymbol{y} = \mathbf{0}$). This alteration would allow us to further distinguish between the effects of spatial and feature-based attention, as feature-based attention should produce a reduction in the baseline firing rate of neurons that are unselective for the presented stimulus feature.

Experimentally, it has been shown that attention-dependent suppression of neural responses is particularly strong when there are multiple stimuli within the cell's RF (Moran and Desimone, 1985; Reynolds et al., 1999). Although we do not explicitly model this effect, it is easy to see how it could come about for our model. When there is one stimulus within a cell's RF, directing attention away or towards the presented stimulus will induce a multiplicative change to the neuron's response, by altering the numerator in equation 4.21. When two stimuli are present within the cell's RF, attending towards one of the stimuli will also alter suppression that comes from the other stimulus, via the denominator in equation 4.21, resulting in larger changes in the neuron's response. Indeed, this effect was demonstrated by Reynolds & Heeger in their normalization model of attention (Reynolds and Heeger, 2009).

4.2.6 Attentional modulation of centre-surround interactions

The responses of neurons in the visual cortex are modulated by stimuli located outside of their classical RF, that do not evoke a response when presented alone. Typically, presenting a stimulus outside of a neuron's RF suppresses its response, compared to when there is only a

single stimulus presented within its RF; a phenomenon called 'surround suppression' (Seriès et al., 2003). Sundberg et al. found that, in visual area V4, attending to a stimulus located within the RF reduces the suppressive influence of a stimulus presented at the surround, while attending to the surround increases this suppression (Sundberg et al., 2009).

We used the setup described in the previous section to measure the degree of surround suppression in our model in the absence of attention, or with attention directed to either the RF centre or the surround (figure 4.11a). By definition, a stimulus in the RF surround should not elicit a response when presented alone, although it may suppress the response of a neuron to a stimulus simultaneously presented in the RF. To reproduce this behaviour in our model we needed to specify the spatial width of the basis functions (λ_{sptl}): if it was too large, surround stimuli would elicit a response when presented alone; too small, and there would be no surround suppression (we found that $\lambda_{sptl} = 2$, produced the required behaviour; figure 4.11a).

Consistent with experiment, directing attention towards the RF increased the model neuron response towards a single stimulus presented within the RF, while decreasing the suppression from a second stimulus presented at the surround (figure 4.11b & c). Directing attention to the surround did not significantly alter the model neuron response when a single stimulus was presented within the RF, but did increase the suppression caused by a second stimulus presented at the surround (figure 4.11b & c). In both attentional conditions, the response of the model neuron to a stimulus presented at the surround alone was negligible.

As before, in order to understand how surround suppression comes about in our model, we consider the probability accorded to different 'explanations' of the sensory input by the agent. When a stimulus is presented at the RF centre alone, there are two likely causes of the resulting sensory input: a stimulus at the RF centre, or no stimulus (with probabilities p_{centre} & p_{none}; figure 4.12a, inner circle). In contrast, the sensory input produced by stimuli at both the RF centre and surround can be accounted for in multiple ways (figure 4.12a, outer circle). Because the agents internal model is assumed to be 'sparse' (i.e. there is a small probability that any particular hidden cause is active) (Olshausen and Field, 2004), the 'true' cause of the sensory input (stimuli at both the RF centre and the surround) is deemed unlikely (p_{both} is small). Instead, they associate equal probability to two alternative explanations: a single stimulus at the RF centre, or the surround ($p_{surr} = p_{centre}$). Overall, presenting a second stimulus at the surround decreases p_{centre}, while only slightly increasing p_{both}, resulting in a reduction in the firing rate of the corresponding model neuron (as $f \propto p_{centre} + p_{both}$).

The effect of attention in our model is to alter the prior probability associated with stimuli at different locations by the agent. In general, the influence of such a perceptual prior will be strongest when there is large uncertainty about the causes of the sensory input. This uncertainty could be due to sensory noise (e.g. at low contrast), or an 'ambiguous' stimulus, where the sensory input is equally likely to be interpreted in more than one way (e.g. the Necker cube

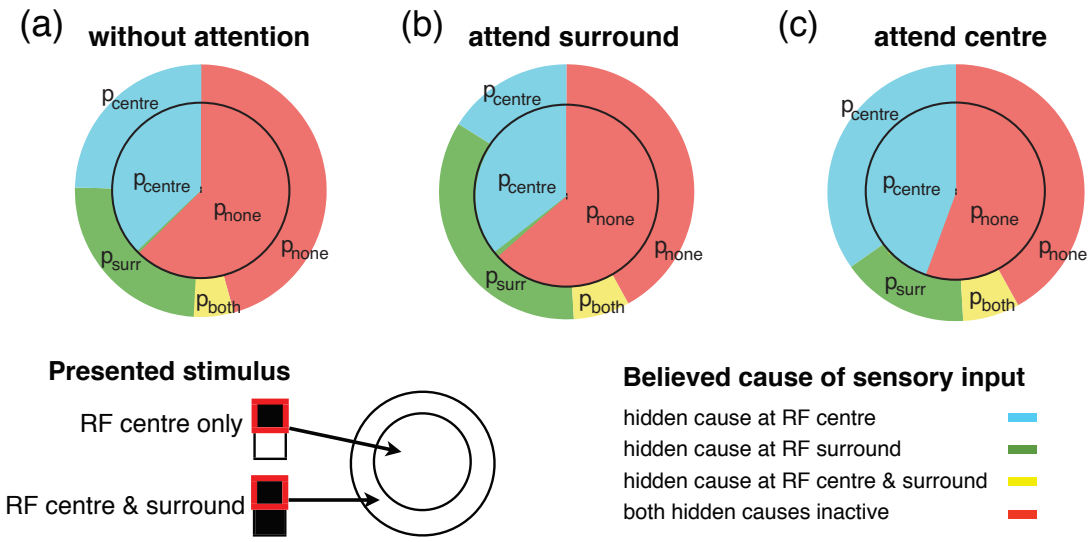


Figure 4.12: Influence of attention on competing explanations of the sensory input, with or without a stimulus presented outside of the RF. Possible explanations are divided into four categories: those where (i) there is an active hidden cause at the RF centre, but not the surround (p_{centre}; blue); (ii) there is an active hidden cause at the RF surround, but not the centre (p_{surr}; green); (iii) hidden causes at the RF centre and surround are both active (p_{both}; yellow); (iv) neither are active (p_{none}; red). (a) In the absence of attention, a stimulus at the surround (outer circle) produces a sensory input that is equally well explained by a single active hidden cause at either the RF centre or surround ($p_{centre} = p_{surround}$). As a result, there is a reduction in p_{centre}, compared to when a stimulus is only present at the RF (inner circle), and a suppression in the firing rate of the corresponding model neuron ($f \propto p_{centre} + p_{both}$). (b) Attending to the surround does not alter the competing explanations when a stimulus is presented at the RF centre alone (inner circle), but has a strong effect when a second stimulus is presented at the surround (outer circle), biasing the agent to interpret the sensory input as due to a hidden cause at the surround (increased p_{surr}, and decreased p_{centre}). This results in an increase in the degree of surround suppression. (c) The converse effect occurs when attention is directed to the centre, resulting in decreased surround suppression.

illusion (Gershman et al., 2009)). In our model, a single stimulus presented at the RF gives rise to an 'unambiguous' sensory input, in the sense that, assuming a stimulus is present, the sensory input can only be explained in one way: due to a single stimulus located at the RF (figure 4.12a, inner circle). As a result, the probability accorded to the different explanations of the sensory input is not strongly altered by spatial attention (figure 4.12b & c, inner circles). In contrast, stimuli presented at both the RF centre and surround give rise to an 'ambiguous' sensory input, which is equally well accounted for by a single stimulus at the RF centre, or the surround (figure 4.12a, outer circle). Consequently, attention-dependent changes to the agents perceptual prior have a large effect in biasing which of these two explanations is preferred, strongly increasing the inferred probability that a stimulus is present at the attended location. Thus, directing attention towards a second stimulus presented at the surround produces a large decrease in p_{centre} (figure 4.12b), resulting in an increase in the degree of surround suppression (as $f \propto p_{centre} + p_{both}$). The converse effect occurs when attention is directed towards the RF centre (figure 4.12c), resulting in decreased surround suppression.

4.2.7 Attention and perceptual transfer

One advantage of learning an internal model that predicts how hidden causes in the world generate the received sensory input, is that this model may be used by the agent to perform many different behavioural tasks. As a result, however, changes in the internal model that take place in order to improve performance in a particular task will alter the agent's performance in other tasks as well.

In our simulations, the agent learned to associate an increased prior probability that hidden causes representing 'target' locations were active (figure 4.5b), resulting in improved detection performance for stimuli presented at these locations (figure 4.6). However, in addition to improving detection performance, this learned prior will also alter the agent's estimation behaviour, so that stimuli are judged as being closer to attended locations than they actually are. This effect is similar to the perceptual estimation biases reported in chapter 3, where participants' learned prior induced an attractive estimation bias towards frequently presented motion directions (figure 3.7). However, in contrast to our psychophysics results, the changes in the agent's prior predicted by our model do not reflect the 'true' stimulus distribution, as stimuli were equally likely to be presented at all locations. Instead, changes to the agent's prior take place in order to compensate for a mismatch between the agent's internal model, in which spatially distributed image features are believed to generate the received reward, and the actual task, in which spatially localized image features determine the received reward. Thus, changes in the perceptual prior that result in improved performance in the detection task may lead to suboptimal perceptual biases and decreased accuracy in an estimation task. We are currently conducting psychophysics experiments to test this prediction.

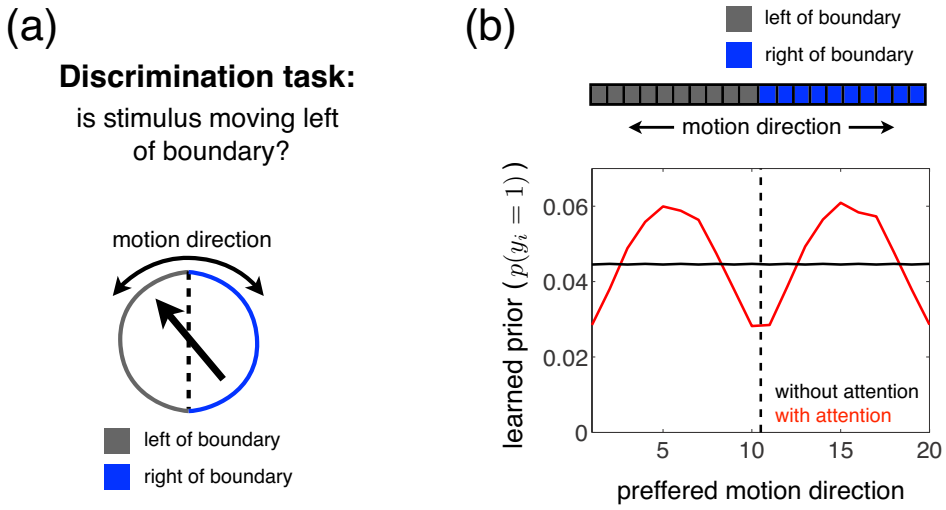


Figure 4.13: Discrimination task. (a) Schematic of task. The agent has to discriminate whether a motion stimulus is moving to the left or the right of a discrimination boundary. (b) After optimization towards the discrimination task, the agent learns to associate an increased prior probability for stimuli moving perpendicular to the discrimination boundaries.

In general, the nature of the perceptual bias that occurs will depend on both the behavioural task and presented stimuli (as well as the structure and form of the agent's internal model; see section 5.1). For example, in a recent psychophysics experiment, Jazayeri et al. found that when subjects are required to discriminate which side of a boundary a stimulus is moving in, they exhibit a repulsive estimation bias away from the discrimination boundary (Jazayeri, 2007).

We adapted our model to investigate the perceptual biases predicted by our model after optimization towards a discrimination (figure 4.13a) rather than a detection task (figure 4.1). Presented stimulus statistics were the same as for our previous simulations (section 4.1.1), with the exception that only one y-unit was present at a time, with equal probability for all units ($p(y_i = 1) = 1/N_y$). The y-unit's in our model represented different stimulus motion directions. Due to the circular basis functions used in our simulations (see section 4.1.2.1), there were effectively two discrimination boundaries (between blue and grey motion directions in figure 4.13a). A binary variable ($t \in \{0,1\}$) denoted whether stimuli were moving to the left or to the right of a discrimination boundaries ($t = 0$ if $\exists \left(1 \leq i < \frac{N_y}{2}\right) : y_i = 1, t = 1$ otherwise). The agent was required to report whether the presented stimulus was moving to the left or right of the discrimination boundary (i.e. whether $t = 0$ or 1), with correct responses followed by an immediate reward ($r = 1$ if $a = t$, $r = 0$ if $a \neq t$).

After optimizing their internal model towards the discrimination task (section 4.1.2.4), the agent learned to associate a decreased prior probability for hidden causes representing stim-

uli moving towards the discrimination boundaries, and increased prior probability for stimuli moving in other directions (figure 4.13b). In our model, this change in the agent's learned prior improves the agent's performance in the discrimination task by increasing the sensitivity of low-level hidden units that are most useful in determining whether stimuli are presented to the left or right of the discrimination boundary. While we did not explicitly model the estimation task, this learned prior will be expected to induce a repulsive estimation bias away from the discrimination boundaries (towards spatial locations with highest prior probability), which is consistent with the experimental results of Jazayeri et al. (Jazayeri, 2007). As with the results of Jazayeri et al., the magnitude of this estimation bias should increase when there is a high degree of perceptual uncertainty, for example with low coherence or low contrast stimuli.

Jazayeri et al. proposed that the repulsive perceptual biases that they observed could be explained by task-dependent changes in the 'decoder', which preferentially weights signals from neurons tuned to motion directions away from the discrimination boundaries. In contrast, in our model, a repulsive perceptual bias would come about due to increases in the gain of low-level visual neurons (e.g. in MT) that are tuned to stimuli moving away from the discrimination boundaries (Scolari and Serences, 2009). While in reality, attention will likely alter neural responses at multiple stages of visual processing, our work cautions against attributing the psychophysically observed perceptual biases as entirely due to changes in the high-level 'decoder', as we show that similar perceptual biases may be produced by changes to the responses of low-level visual neurons.

Rather than providing a quantitative fit to the results of Jazayeri et al., our aim was to demonstrate in principle, how qualitatively different perceptual biases are produced by different behavioural tasks (e.g. attractive biases for a detection task, repulsive biases for a discrimination task). We show that, contrary to the standard view of visual attention, which is usually thought to select a particular spatial location or visual feature for increased processing, task-dependent changes to visual processing can be complex; and very far from the typical 'spotlight' analogy of visual attention.

4.3 Discussion

An important goal in visual neuroscience is to understand why the response properties of visual neurons are the way they are. We extended previous statistical models of visual processing (Hyvärinen, 2010) to account for the effect of behavioural demands on visual neuron responses, hypothesizing that the brain learns a probabilistic model that predicts how both the sensory input and reward received for performing different actions are determined by a common set of hidden causes (Sahani, 2004). This framework has two main advantages. First, it has predictive power: in theory, changes to neural responses should be predicted as a di-

rect consequence of the presented stimuli and behavioural task. Second, predicted changes to neural responses have a direct functional meaning: they correspond to changes in the believed causes of the sensory input.

To predict how goal-directed attention should modulate neural responses, we needed to make certain assumptions. First we assumed that the firing rate of each neuron encodes the probability that a particular hidden cause contributed to generating the received sensory input. Second we assumed that the agent learns a 'sparse' model, with a small prior probability that any particular hidden cause is active (Olshausen and Field, 2004). This was required to produce surround suppression in our model, which came about due to competition between different possible causes of the sensory input. Third we assumed that the internal model is hierarchical (Karklin and Lewicki, 2005), with high-level hidden causes assumed to be responsible for generating the received reward. Finally, we assumed that attention alters the sensitivity of individual neurons (by changing the prior probability that hidden causes are active), but not the network connectivity (the basis functions). These assumptions are not new, but on the contrary, are often included in phenomenological and mechanistic models of attention (Reynolds and Heeger, 2009; Ghose, 2009; Lee and Maunsell, 2009). However, in contrast to these models, we justify our assumptions from functional principles, in order to provide insight into 'why' goal-directed attention alters visual neuron responses as it does.

Our model predicts a range of task-dependent changes to neural responses that are qualitatively consistent with experimental observations in low to mid-level areas of the visual cortex: modulation of neural contrast response functions (figure 4.7), sensory tuning curves (figure 4.10) and centre-surround suppression (figure 4.11). Both the predictions and mathematical formulation of our model bear strong similarities to the 'normalization model of attention', proposed by Reynolds & Heeger (Reynolds and Heeger, 2009). In common with our work, Chikkerur et al. recently showed that Reynolds & Heeger's model can be derived using a Bayesian framework (Chikkerur et al., 2010) (section 2.3.2). However, while Chikkerur et al. explicitly specified an ad hoc attentional prior, in our work, task-dependent changes to the internal model are learned automatically, to improve predictions of the received reward.

Several studies have tried to explain visual attention in Bayesian terms, under the hypothesis that it corresponds to changes in the perceptual prior (Dayan and Zemel, 1999; Rao, 2005; Chikkerur et al., 2010; Yu and Dayan, 2005b; Yu et al., 2009). However, in these studies, attention-dependent changes to the prior were either specified explicitly (Dayan and Zemel, 1999; Rao, 2005; Chikkerur et al., 2010), or learned from the statistics of stimuli presented during the task (Yu and Dayan, 2005b; Yu et al., 2009). In contrast, we investigate how visual processing is influenced by behavioural demands, in the absence of any changes to the presented stimulus statistics. Within our proposed framework, behavioural demands alter visual processing when there is a mismatch between the internal model and the external environment.

We hypothesize that such a mismatch occurs when task-relevant stimuli are more localized than the high-level features in the agent's internal model. In this case, attention will alter visual processing to improve predictions of the received reward, at the possible expense of learning a worse model of the sensory inputs.

In addition to improvements in behavioural performance, we predict that attention-dependent changes to the learned prior should alter the agent's performance in other tasks, giving rise to estimation biases such as those observed in our psychophysics experiment (chapter 3). However, in contrast to the estimation biases that we observed in our experiment, our model predicts that estimation biases should be induced by changes to the behavioural task alone, in the absence of any changes to the presented stimulus statistics. A similar effect could underlie longer term perceptual biases that are observed experimentally. For example, the owl is systematically biased to estimate sounds as coming from directions closer to their centre of gaze than they actually are. Fischer et al. showed that this perceptual bias is consistent with Bayesian estimation, with a prior that favours central directions (Fischer and Peña, 2011). However, as this prior is defined relative to the position of the owl, it cannot correspond to the true distribution of sound directions in the world, but instead must represent the 'relevance' of the different directions.

To fit the joint distribution over the reward and sensory input, model parameters should be learned to maximize the objective function:

$$L(\theta,\psi) \quad = \quad \sum_i \left[\log p\left(r_i|\boldsymbol{x}_i,a_i;\theta,\psi\right)+\log p\left(\boldsymbol{x}_i|\theta\right)\right]. \tag{4.24}$$

As the sensory input ($\boldsymbol{x}_i$) will typically have many more dimensions than the received reward, the second term of equation 4.24 will usually dominate learning. In line with this, we found that, when models parameters were learned in order to maximize equation 4.24, the received reward had a negligible influence on the internal model that was learned.

Previous work has mostly focussed on the case where the internal model model is learned in order to maximize the log-probability of the received sensory input (equivalent to maximizing the second term of equation 4.24). In an exception to this, Sahani (2004) considered representational learning with an objective function similar to equation 4.24 (Sahani, 2004). Sahani postulated reward-dependent 'weighting terms' that determine the relative magnitude of the first and second terms of the objective function (although he did not specify how these weightings should depend on the reward). Sahani's work provides an interesting framework for considering how the behavioural relevance of the received sensory input could influence representational learning. However, Sahani did not investigate the implications of the framework for sensory neural responses. Our work can be seen as an extension to Sahani's; we show that, under certain assumptions about the internal model in the brain, the framework can be used to make experimentally testable predictions about the affect of visual attention on neural responses.

We were interested in how behavioural demands influence sensory processing, rather than the presented stimulus statistics *per se*. Therefore, we studied how the internal model adapts in order to improve predictions of the received reward (by maximizing the first term of equation 4.24), regardless of how well it predicts the received sensory input (the second term of equation 4.24). However, in general, the agent will have to achieve a balance between optimizing the internal model towards current task-demands, and learning a good representation of the sensory inputs that allows generalization across different tasks. We discuss this further in section 5.2.

In our work, the agent learned to predict the probability distribution over the received reward, given the performed action and received sensory input ($p(r|x,a;\theta,\psi)$). However, they only used the mean of this distribution ($\langle r \rangle_{p(r|x,a;\theta,\psi)}$) to choose which action to perform. Thus, the agent learned information about the high-order reward statistics that was not required to perform the task, which could be inefficient. In comparison, in most reinforcement learning algorithms the agent learns to predict the mean reward associated with each action, while neglecting higher order reward statistics (Sutton and Barto, 1998). However, there are certain reasons why it may be advantageous for the agent to learn about more than just the expected reward. First, the agent's belief about the higher-order reward statistics will play a role in determining how the internal model is learned. For example, the received reward will have a much stronger influence in driving changes in the internal model if the predicted reward distribution is very narrow. Second, the agent may want to optimize more than just the expected reward; to avoid bankruptcy, a poker player might seek to limit the variance in their takings, in addition to maximizing their long-term gains.

At present, it is unknown how (or indeed, whether) probability distributions are represented in the brain (Fiser et al., 2010; Shelton et al., 2011; Deneve, 2008a; Ma et al., 2006). A current area of debate is whether neural firing rates encode samples from a probability distribution (Fiser et al., 2010; Shelton et al., 2011); or parameters, such as the mean and variance of the distribution (Deneve, 2008a; Ma et al., 2006). In our simulations, we assumed that mean firing rates are proportional to the probability that individual hidden causes contributed to generating the received sensory input. While this coding scheme was chosen for simplicity, it produces means firing rates that are qualitatively consistent with a 'sampling' code (Shelton et al., 2011). Meanwhile, certain parametric codes, such as the coding scheme proposed by Deneve et al. (Deneve, 2008a), predict mean firing rates that are qualitatively consistent with our model (i.e. they scale monotonically with the posterior probability that encoded latent variables are 'active').

We investigated short term effects of behavioural context, focussing specifically on visual attention. We hypothesized that over these timescales, only the sensitivity of individual neurons (the prior) varies, while the network connectivity (the basis functions) remains constant.

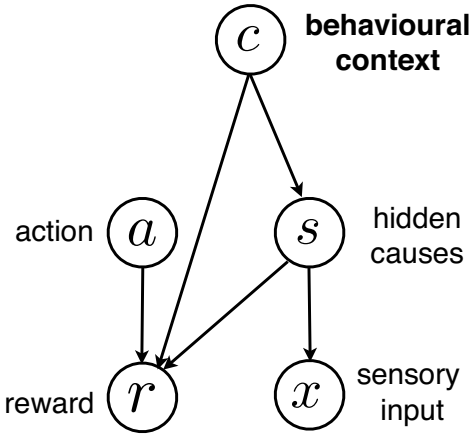


Figure 4.14: Hypothetical internal model, in which a latent variable (c) denotes the current behavioural context, which determines the prior distribution over hidden causes ($p(s|c)$) and the agent's model of the task ($p(r|a,s,c)$). The agent can use their received sensory input and reward to *infer* the current behavioural context, enabling the focus of attention (determined by $p(s|c)$) to be shifted more rapidly than if they had to learn the task from scratch.

This restriction could be removed to investigate changes that take place over longer timescales. Currently, the relationship between different types of sensory learning (e.g. 'attentional' (Eckstein et al., 2004; Jiang and Chun, 2001) versus 'perceptual' learning (Fahle, 2005; Seitz et al., 2009)), and how they depend on the training paradigm, is an active area of research. Hopefully, our framework could contribute towards this debate.

In our work attention needs to be learned; it requires optimizing the response properties of sensory neurons based on feedback in a task. This is a good description of what must happen when an animal is presented with a novel task. However, attention can also be directed quickly, without requiring task-feedback. To account for this, we could extend our model by including additional latent variables in the agent's internal model which represent different behavioural contexts. Thus, the agent could use their received sensory input and/or task-feedback input to '*infer*' that appropriate behavioural context, allowing them to direct their attention more quickly than if they had to learn the behavioural context from scratch (figure 4.14).

We put forward a very general framework for predicting how task-demands should alter visual processing. We then showed that, given certain assumptions about the internal model and behavioural task, this framework predicts attention-dependent changes to neural responses that are consistent with existing phenomenological models of attention. However, although the assumptions of our model are based on functional principles, in order to truly 'derive' the effects of attention it would be desirable construct a more sophisticated model of natural images, in which the model parameters are learned directly from natural image statistics (as opposed to artificial data). In the past, this approach has been highly successful in understanding the

passive properties of visual neurons. Hopefully, in the future, it could be used to make quantitative and testable predictions about how different behavioural tasks alter visual processing and perception.

Chapter 5

Discussion

In this chapter, we discuss three broad issues that have important implications for our work: the structure and form of the internal model, how the internal model should be updated to deal with new information about the environment, and the neural implementation of Bayesian inference. We then state the main conclusions of the thesis, and discuss the relevance of our work to the questions posed in our literature review (chapter 2): why are goal-orientated attention and expectations necessary, how are they controlled, and what is their effect on visual processing and perception?

5.1 Structure and form of the internal model

What information is encoded by sensory neurons? A normative approach to this question involves asking what information 'should' encoded if visual processing were adapted towards the environment. Thus, rather than studying sensory processing directly, the researcher begins by constructing a probabilistic model that is able to capture the statistical structure of natural images. Ideally, the model can then be used to make neurophysiological or perceptual predictions, to be tested experimentally.

However, due to the complexity of real-world environments, the visual system is unlikely to learn a perfect model of how hidden causes in the world generate the received sensory input; in most cases there will be some mismatch between the internal model, and the true structure of the environment. In chapter 4 we argued that, faced with such a mismatch, the visual system should prioritize aspects of the model that are important in determining behaviour. Under this assumption, we investigated how the internal model should adapt in response to changing task-demands.

More generally however, behavioural demands may also play a role in determining how the internal model is learned in the first place, during evolution and development. For example, the visual system could allocate a disproportionate number of neurons to encode behaviourally

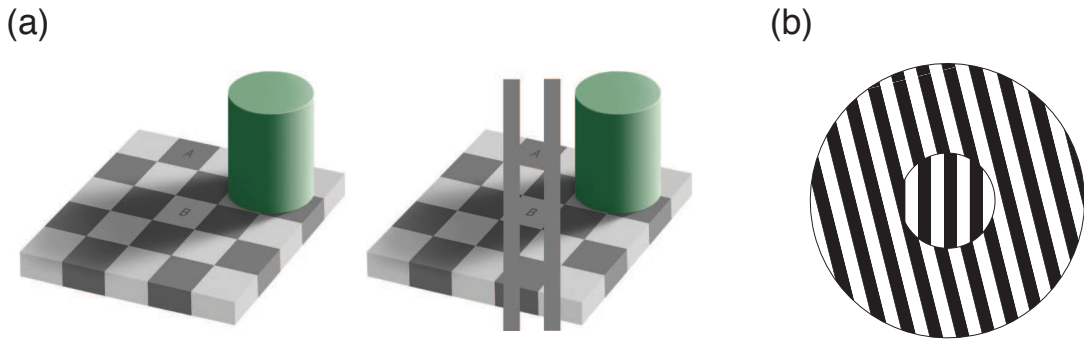


Figure 5.1: (a) Checker-board illusion. In the left panel, the square marked *'B'* appears to be lighter than the square marked *'A*'. The right panel shows that *A* and *B* are in fact identical. (b) Tilt illusion. The presence of the surround stimulus, tilted 15° anti-clockwise from vertical causes the central (vertical) stimulus to appear tilted clockwise. It has been proposed that the underlying functional explanation for this tilt illusion is very similar to the checker-board illusion.

relevant sensory dimensions. This idea is supported experimentally: in one of the few experimental tests of the 'efficient coding hypothesis', Machens et al. found that grasshopper auditory neurons are optimized to encode behaviourally relevant sounds, such as communication signals, rather than the sounds that are found in their natural habitat (Machens et al., 2005). Thus, an interesting direction for future research would be to investigate how the nervous system is optimized towards 'natural reward statistics', as well as sensory input statistics (Montague and King-Casas, 2007).

5.1.1 Dependence of perceptual biases on the internal model

The form of the internal representation will determine how perception of visual stimuli is influenced by their statistical context. For example, consider the classic checker-board illusion shown in figure 5.1. In this illusion, a square that is located within a shaded region of the image (*B*) is perceived as being brighter than a square located at an illuminated region of the image (*A*), despite the fact that they are both identical. If the goal of visual processing were to infer the absolute *brightness* at each point in the image, then this illusion would be paradoxical; if anything, the fact that neighbouring pixels are likely to be of a similar brightness should cause the square located within the shaded region to appear darker than it actually is. However, the illusion can be explained if we assume that the visual system takes into account the level of *illumination* (a global property, that will tend to be statistically coordinated for nearby regions of the image), to infer the *surface reflectance* at each point in the image (typically, a local property). Thus, in figure 5.1, a square located in a shaded region of the image is judged as having a higher reflectance than another square, of identical greyscale, located in an illuminated region of the image. That is, what seems like an incorrect judgement about brightness (or

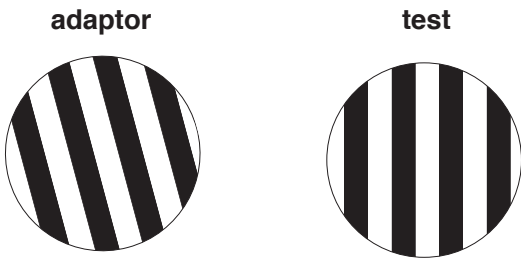


Figure 5.2: Tilt after-effect. After looking at the grating on the left for at least 30 seconds, the grating on the right should appear tilted clockwise.

greyscale), in fact corresponds to correct inference about the surface reflectance of the objects that generated the image.

A similar argument was put forward by Schwartz et al. to explain the tilt illusion, where a central orientated stimulus is perceived to be tilted *away* from its surrounding visual context (figure 5.1b) (Clifford et al., 2000; Schwartz et al., 2009). If the goal of visual processing were to infer the local orientation at each point in the visual scene, then the surrounding visual context should have the opposite effect; the central stimulus should appear tilted *towards* the surrounding context. Instead, Schwartz et al. proposed that the visual system learns a generative model that describes how the local image structure (analogous to reflectance) is combined with the global image structure (analogous to illumination) to generate the received sensory input (Simoncelli and Schwartz, 1999; Schwartz and Simoncelli, 2001; Schwartz et al., 2006). The global structure does not have a direct physical interpretation, but corresponds to features such as orientated textures and edges, which cause the orientation of nearby locations to be statistically correlated. Thus, perceptual biases such as the tilt-illusion come about because the visual system takes the global image structure into account (dictated by the orientated surround in figure 5.1b) to infer the local structure at each point in the image. Analogous to the checker-board illusion, incorrect judgements about the orientation of the central stimulus in figure 5.1b correspond to correct inferences about its local structure (i.e. the *difference* in orientation between the centre and surround).

There are strong similarities between the perceptual biases produced by the temporal and spatial context of visual stimuli (e.g. compare the 'tilt after-effect illusion' shown in figure 5.2 with the 'tilt-illusion' shown in figure 5.1b). Consequently, it has been suggested that similar functional principles could underlie both types of perceptual bias (Schwartz et al., 2007). For example, the model of Schwartz et al. (2009) can be adapted to explain how changes in temporal context give rise to the tilt after-effect illusion (figure 5.2) (Wainwright et al., 2001). First, assume that people learn a generative model describing how local *temporal* structure combines with global *temporal* structure, to generate the received sensory inputs. The global tempo-

ral structure would correspond to temporal correlations in sensory signals. Analogous to the tilt-illusion, perceptual illusions would come about because the visual system takes the global temporal structure into account, in order to infer the local temporal structure at each moment in time. Thus, incorrect judgements about the orientation of the test stimulus in the tilt after-effect illusion (figure 5.2) would correspond to correct judgements about its local temporal structure (i.e. subjects infer the *difference* in orientation between the adaptor and the test stimulus).

The issues raised by these studies are relevant to both our psychophysics and modeling work. In our psychophysics experiment, we modelled subjects' behaviour by assuming that they performed perceptual inferences about the stimulus motion direction. Under this assumption, our model predicted attractive estimation biases towards frequently presented motion directions, similar to what was observed experimentally. However, previous experiments have reported that the qualitative effects of spatiotemporal context on perceived stimulus orientation vary with contrast: attractive estimation biases are observed at low contrasts, while repulsive estimation biases are observed at high contrasts (Roberts and Thiele, 2008). Our simple Bayesian model can not account for the repulsive biases that occur at high contrast. Thus, understanding how spatiotemporal expectations alter the perception of high contrast, as well as low contrast stimuli, may require a richer description of the subject's internal representation; informed by the statistical structure of natural images.

In our modeling work, the agent learned to associate an increased prior probability for behaviourally relevant stimuli, leading to the prediction that people should exhibit an attractive estimation bias towards attended stimuli. However, as discussed, the perceptual bias that is predicted will depend on the form of the internal model; if instead, we were to assume that the agent learned a more complex internal model, describing how the local image structure is combined with global image structure, we might expect similar changes in the prior to produce qualitatively different perceptual biases, *away* from the attended stimulus.

5.1.2 Influence of internal model structure on generalization & specificity of learned expectations

The form of the internal model will also determine how subjects' prior belief about the likely motion directions generalize to alter their perception of stimuli that differ in one or more sensory dimensions (e.g. colour, or speed). For example, if subjects use an internal model in which 'colour' and 'motion direction' are assumed to be independent ($p(\text{colour}, \text{direction}) = p(\text{colour})\,p(\text{direction})$), then their learned expectations over motion direction will give rise to biases in the perceived motion direction that are independent of the stimulus colour. Conversely, if subjects use an internal model where 'colour' and 'motion direction' are represented jointly ($p(\text{colour}, \text{direction})$), then their learned expectations over motion direction will give rise to perceptual biases that depend on the stimulus colour.

Gekas et al. adapted our experiment to investigate the specificity of stimulus expectations acquired through exposure to distinctly different motion direction distributions, differentiated by colour (Gekas et al., 2011). Presented stimuli could be either red or green: green stimuli were presented most frequently moving in one of two directions; red stimuli were presented with equal frequency moving in all directions. Consistent with our experimental results, Gekas et al. found that subjects learned to expect the two most frequent motion directions, with attractive estimation biases towards these directions, as well as hallucinations when no stimulus was presented. Interestingly, subjects exhibited similar estimation biases for both stimulus colours, despite the fact that the distribution of presented motion directions was different for each colour. Further, subjects' estimation distributions for trials when no stimulus was presented did not depend on whether they reported seeing a red or a green stimulus. These results led Gekas et al. to conclude that subjects learned a single prior distribution over motion direction, without taking into account the differences between the stimulus distributions for each colour.

Subjects' inability to learn that the distribution of presented motion directions depended on stimulus colour led them to perform suboptimally in the psychophysics task of Gekas et al. A possible normative explanation for this behaviour comes from the fact that in most real-world situations, stimulus colour is unrelated to motion direction. Thus, while subjects' prior beliefs for colour to be independent of motion direction leads to suboptimal performance in tasks involving artificial stimulus statistics, it could reflect optimal adaptation towards the statistics of the environment. However, this normative explanation seems to be contradicted by a well known illusion called the 'McCullough effect', where the perceived colour of a black and white grating is altered after a brief exposure to a coloured grating with a similar orientation (McCollough, 1965). This interaction between perceived colour and orientation suggests that people learn a joint representation of colour and orientation. It is difficult to justify this representation on normative grounds: why would subjects' learn that motion direction and colour are independent, but not orientation and colour? A more parsimonious for these effects can be made on anatomical grounds: while neurons in the ventral stream encode both colour and orientation (Johnson et al., 2008), it has been postulated that motion direction and colour are processed separately, within the dorsal and ventral pathways respectively (Mishkin et al., 1983) (although see (Thiele et al., 2001)). Indeed, this discussion highlights a limitation in the normative approach: to provide a complete account of visual processing, ideal observer models should be combined with algorithmic constraints, such as anatomical or metabolic constraints faced by the visual system (Marr, 1982).

5.2 How is the internal model altered by experience?

In this thesis, we investigate how visual processing adapts in response to changes in the statistics and behavioural relevance of received sensory inputs. However, in both our experimental and theoretical work, we focus on characterizing 'what' is learned, rather than the learning process itself. In our psychophysics experiment, the rapid speed with which people learned which stimuli to expect prohibited us from measuring its time-course. In our modeling work, we analyzed the predicted changes to visual neuron responses following optimization in a behavioural task, but did not study how this attentional modulation is 'learned' when the agent is presented with a novel task (Chun, 2000; Dayan et al., 2000). Thus, an obvious extension to our work would be to investigate the process by which prior beliefs are learned from sensory experience and task-feedback.

5.2.1 Frequentist versus Bayesian learning algorithms

In a frequentist 'maximum-likelihood' (ML) learning algorithm, internal model parameters (θ) are learned in order to maximize the objective function:

$$L(\theta) = \sum_{i=1}^{N} \log p(x_i|\theta), \tag{5.1}$$

where the summation is taken over a bock of N trials and x_i denotes the sensory input or reward received on the i^{th} trial. For an online learning algorithm, model parameters could be updated after each trial, according to:

$$\theta_{new} = \theta + \eta \partial_\theta \log p(x_i|\theta), \tag{5.2}$$

where η denotes the learning rate. With a suitable choice of step-size (η), θ should converge on a local maximum of $L(\theta)$ after many trials.

An alternative, Bayesian, learning algorithm requires that the subject represents the posterior probability distribution over θ, given the received sensory input: $p(\theta|\boldsymbol{X}_t)$ (where $\boldsymbol{X}_t = (x_1, \ldots, x_t)$ denotes all the inputs up until time t). On receiving a new sensory input (x_{t+1}), the ideal observer should adapt their belief about the internal model parameters according to:

$$p(\theta|\boldsymbol{X}_{t+1}) = p(x_{t+1}|\theta)\, p(\theta|\boldsymbol{X}_t). \tag{5.3}$$

Most previous statistical models of visual processing assume that the agent learns a single set of internal model parameters, using a maximum likelihood algorithm (Hyvärinen, 2010). To facilitate comparison with this body of work, we used a similar modelling framework; we assumed that the agent learns a single set of internal model parameters, updated online according to equation 5.2.

When there is a large set of training data (we used 10^5 trials in our simulations), the ML and Bayesian learning algorithms will usually give very similar predictions (i.e. $p(\theta|\boldsymbol{X}_t) \rightarrow \delta(\theta - \theta_{ML})$ as $t \rightarrow \infty$). However, with a limited supply of unreliable training data, this will not be the case. Further, both learning algorithms make different predictions about how the agent should update their prior belief on each trial. For the ML learning algorithm, the learning rate is determined by a free parameter, η. In contrast, for the Bayesian learning algorithm, the learning rate is constrained by the received sensory input itself. With the Bayesian learning algorithm, the learning rate will depend on the agent's prior uncertainty about the internal model parameters, as well as their prior beliefs about how fast the world changes (see section 5.2.3).. Thus, when the agent is faced with a new environment and receives 'unexpected' sensory inputs (i.e. $p(x_t) = \int p(x_t|\theta)\, p(\theta|\boldsymbol{X}_{t-1})\, d\theta$ is small), they will learn to associate a greater degree of uncertainty over θ. As a result, new sensory inputs will have a stronger influence in altering their prior beliefs, giving rise to fast learning. As they learn the statistics of their new environment, the uncertainty in their estimate of the internal model parameters will decrease, with a corresponding decrease in the learning rate.

Future work investigating how prior beliefs are learned from experience could compare the predictions of both learning algorithms, with the aim of making predictions that can be tested psychophysically (Yu and Dayan, 2005b; Yu et al., 2009). As well as improving our understanding of the learning process itself, this work could be used to address deeper questions about how people use and structure environmental knowledge. That is, to what extent do people represent uncertainty about the 'structure' of their environment (i.e. model parameters/model class), in addition to uncertainty about the stimulus features (i.e. hidden variables)?

5.2.2 Psychophysical measurement of learning dynamics

In our psychophysics experiment, subjects learned to expect the likely stimuli within very few trials (figure 3.6), making it difficult for us to measure the short term time-scale and dynamics of learning. One way around this would be to use a more complex stimulus distribution, so that learning occurred more slowly. Alternatively, we could alter our experiment by repeatedly changing the stimulus distribution. For example, after a fixed number of trials, we could alter the frequently presented motion directions, so that participants had to relearn the stimulus distribution. As a result, we would be able to collect more data from trials in which rapid learning took place, allowing us to measure its dynamics and time-course more accurately. However, we would have to be careful to consider the effect of subjects' previously learned expectations, which would influence how they learned the new stimulus distribution.

Eckstein et al. conducted a psychophysics experiment to investigate how humans alter their prior expectations based on feedback in a task (Eckstein et al., 2004). In their experiment, subjects had to localize a target stimulus of unknown identity, presented at one of several locations.

After each trial, they were provided with feedback about the location of the target, which they could use to update their prior belief about its identity. Learning took place over blocks of 4 trials, after which the target identity was changed. Eckstein et al. compared subjects' performance in the task to an ideal Bayesian observer, who updated their prior beliefs according to equation 5.3. Similar to our results, Eckstein et al. observed rapid learning over blocks of only 4 trials. However, they found that human learning was slower than would be predicted for an ideal Bayesian observer, with subjects relying more heavily on their previous decisions in the task than the provided feedback, and with virtually no learning on trials following a previous incorrect response. This work shows how the concept of an ideal Bayesian observer can be used to give insight into human perceptual learning. While Eckstein et al. looked exclusively at changes in perceptual performance, their experiment could be adapted to investigate how perceptual biases, such as those observed in our experiment, are acquired.

5.2.3 Influence of learning algorithm on perceptual biases that develop over different timescales

In chapter 3 we showed that subjects learn to expect frequently presented stimulus features, resulting in attractive estimation biases towards these features (figure 3.7). However, under different circumstances, changes in the presented stimulus statistics have been found to produce qualitatively different types of perceptual bias. For example, prolonged exposure to a visual pattern, such as an orientated grating, can result in sensory *adaptation*, where subsequently presented stimuli are perceived as being more dissimilar to the original ('adaptor') stimulus than they actually are (Levinson and Sekuler, 1976; Schwartz et al., 2007; P Seriès and Simoncelli, 2008) (figure 5.2). Perceptual adaptation is usually accompanied by a reduction in the sensitivity of visual neurons that are tuned to the adaptor stimulus (Carandini, 2000). Similar effects have also been observed over longer time scales. For example, after several hours of looking through a camera that selectively filters out a specific orientation, subjects exhibit increased perceptual sensitivity for the filtered orientation (Zhang et al., 2009). This contrasts with our results, where subjects exhibited increased perceptual sensitivity for motion directions that were observed more frequently (figure 3.10).

The perceptual biases observed in our experiment were accounted for by assuming that subjects learned to associate an increased prior probability for frequently presented stimuli (figure 3.18). However, a similar short-term change in the perceptual prior cannot account for the repulsive biases that are observed as a result of sensory adaptation (although see section 5.1.1). Instead, it is possible that these repulsive biases are associated with a change in the observer's likelihood function, that predicts how hidden causes in the world generate the received sensory input (Buiatti and van Vreeswijk, 2003; Stocker and Simoncelli, 2006b; Schwartz et al., 2007). For example, a subject looking through a camera that filters out a particular orientation might

(correctly) deduce that there has been a change in how their sensory inputs are generated, such that a stimulus presented at the filtered orientation gives rise to a weaker sensory signal than the same stimulus presented at some other orientation. To compensate for this change, the subject could alter their likelihood function, so that a weak sensory signal at the filtered orientation was believed to be generated by the same hidden cause as a stronger sensory signal at some other orientation. While this change in the subject's likelihood function would help to optimize visual processing while they were looking through the camera, it would lead to suboptimal perceptual biases when the camera was removed.

Under what conditions should people exhibit attractive biases associated with 'expectations', as opposed to repulsive biases associated with 'adaptation'? While numerous examples of both attractive and repulsive perceptual biases are found in the experimental literature, we are far from having a clear answer to this question. One factor that could be important is the timescale over which changes in the stimulus statistics occur. For example, if the level of illumination, which determines how visual signals are produced by hidden causes in the world, typically varies more quickly than the distribution of hidden causes, then the ideal observer should attribute short-term changes in the stimulus statistics to a change in the level of illumination, and longer-term changes in the stimulus statistics to changes in the prior distribution over the hidden causes. In this case, over short timescales the ideal observer would alter their likelihood function, giving rise to repulsive perceptual biases, while over longer timescales they would alter their prior, giving rise to attractive perceptual biases. Experimental support for this idea was provided by Kanai & Verstraten, who found that varying the time interval between an adaptor and a test motion stimulus caused a reversal in the direction of subjects' estimation biases (Kanai and Verstraten, 2005). When there was only a brief time interval between the adaptor and test stimulus, subjects exhibited a repulsive bias, and were more likely perceive the test stimulus moving in the opposite direction to the adaptor. With a longer time interval, subjects exhibited an attractive bias, and were more likely perceive the test stimulus moving in the opposite direction to the adaptor.

In general, whether the agent should alter their likelihood function over shorter or longer timescales than their prior, will depend on the temporal structure of the environment. Thus, under the assumption that the visual system is optimized towards the natural environment, the speed at which different aspects of the internal model vary in time in response to changes in the presented stimulus statistics will be determined by the temporal structure of natural sensory signals. Consequently, future work studying the temporal statistics of natural movies or image sequences, could be used to predict the changes in the agent's internal model, and thus, the perceptual biases, that occur over different timescales.

5.2.4 Reward-driven learning: relation to reinforcement learning & decision theory

In chapter 4 we posited that the purpose of visual processing is to represent sensory information in a way that facilitates interaction with the environment. Previous work has hypothesized that the visual system achieves this goal by learning a probabilistic model that predicts how hidden causes in the world generate the received sensory input (Olshausen and Field, 1996; Hyvärinen, 2010). Visual processing would then consist of inferring the hidden causes that generated a given sensory input. An implicit assumption underlying this work is that the hidden causes of natural sensory signals are of direct behavioural relevance to the organism. Thus, the visual system is assumed to take advantage of the inherent structure of natural sensory signals to learn a sensory representation that is useful for performing many different behavioural tasks (Gershman and Niv, 2010).

However, in some cases the image features that are relevant to a particular task may differ from the hidden causes learned by the agent. We hypothesized that the visual system deals with this eventuality by adapting the internal model to optimize performance in the current task, at the possible expense of learning a worse model of the received sensory inputs. That is,– we assumed that over short timescales the internal model (parameterized by $\{\theta, \psi\}$; see section 4.1.2.4) adapts to improve its predictions for the received reward ($p(r|x,a;\theta,\psi)$), regardless of how well it predicts the received sensory inputs ($p(x;\theta)$). However, this learning procedure has certain weaknesses. First, the behavioural task will typically be highly labile and provide limited constraints on the internal model, compared to the bulk of the visual input. Second, optimizing the internal model towards a particular behavioural task will likely reduce its ability to generalize across different tasks. It is possible that these problems are reflected in human behaviour: for example, repeated practice in a behavioural task can lead to 'negative transfer', where performance in a different task is reduced (Seitz et al., 2005a) (see section 4.2.7). Further work will be required to understand how the visual system combines information from sensory signals and task-feedback to adapt towards current behavioural demands, while maintaining its ability to generalize across different behavioural contexts.

A number of researchers have used ideas from reinforcement learning and decision theory to investigate how reward is represented in the brain, and how decisions are made in the face of uncertainty (Schultz et al., 1997; Doya, 2008; Dayan, 2008; Rangel et al., 2008; Schultz, 2006). However, relatively little work has studied how the reinforcement task influences the sensory representation itself, which is usually treated as a fixed input to the model. To address this question,– we chose to start from a standard theoretical account of unsupervised representational learning, which has been used previously to study the response properties of visual neurons (Sahani, 2004; Hyvärinen, 2010). In the future, it would be interesting to investigate how unsupervised learning algorithms can be combined with a reinforcement learning

framework, to optimize interaction with the environment. For example, in complex real-world environments, traditional reinforcement learning algorithms often behave very badly, and thus it may be advantageous for the agent to use prior knowledge about the statistical structure of real-world tasks and sensory signals to help simplify learning (Courville et al., 2006; Gershman and Niv, 2010).

5.3 Neural implementation of Bayesian inference

A large body of behavioural evidence suggests that humans take uncertainty into account when making perceptual inferences about the world (Knill and Richards, 1996b; Rao et al., 2002). As a result, researchers have begun to ask how Bayesian inference might be implemented in the brain. This research addresses two basic questions: how are probability distributions represented by populations of spiking neurons, and how do neural circuits implement probabilistic inference and learning with these representations? While a number of probabilistic neural codes have been proposed, at present there is little consensus over which of these is the most plausible, or indeed, better supported by experimental data. Here, we give a brief overview of the main classes of model, explaining how they relate to our work.

5.3.1 Probabilistic population coding

A conventional view of neural coding is that a population of neurons encodes information about a stimulus, s, through their firing rates, $\boldsymbol{r}$. Thus, the presented stimulus value can be estimated from the population response ($\hat{s} = g(\boldsymbol{r})$) (Seung and Sompolinsky, 1993; Pouget et al., 1998). In contrast, *probabilistic population codes* hypothesize that the neural population response ($\boldsymbol{r}$) encodes the posterior probability distribution over possible stimulus values, given the received sensory input ($p(s|x)$).

A well-known example of a probabilistic population code is 'gain encoding' (Ma et al., 2006, 2008). In this coding scheme, individual neurons are assumed to encode a stimulus s with a mean firing rate $f_i(s)$ and Poisson-like noise. With homogeneous bell-shaped tuning-curves, a given stimulus will produce an activity profile (a plot of neural firing rates versus preferred stimulus value) that has a single peak close to the neuron that is tuned to the presented stimulus. 'Gain encoding' posits that the mean of the posterior distribution is encoded by a weighted sum of the neural activities (roughly related to the peak of the activity profile), while the variance of the distribution is related to the magnitude of the activity profile; larger responses would correspond to a smaller variance (figure 5.3). Pouget and colleagues have shown that this coding scheme can be used to provide a simple implementation of probabilistic computations, such as cue combination (Ma et al., 2006) and optimal decision making (Beck et al., 2008). However, this and related work (Zemel et al., 1998; Jazayeri and Movshon, 2006; Ma et al.,

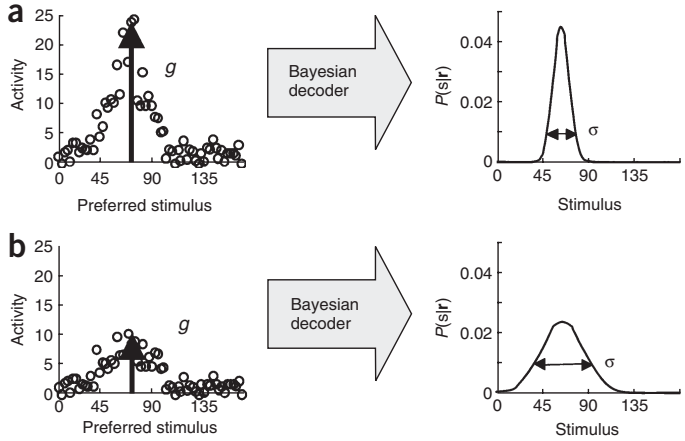


Figure 5.3: Schematic of a 'gain encoding' scheme (adapted from (Ma et al., 2006)). (a) The left panel shows the population response, $\boldsymbol{r}$, to a stimulus whose value is $s = 20$. Neural firing rates are plotted as a function of their preferred stimulus value (i.e. the stimulus that corresponds to the peak of their tuning curve). The right panel shows the posterior probability distribution over s, decoded from the population response using Bayes' theorem. With Poisson-like neural variability, the variance of the posterior distribution (σ^2) will be inversely proportional to the amplitude of the population response (g). (b) Decreasing the amplitude of the population response increases the width of the encoded distribution.

2006) has mostly studied how neural populations encode simple low-dimensional stimulus features (although see (Sahani and Dayan, 2003) for an exception). Thus, it is not clear how these codes scale to encode high-dimensional stimuli such as those used in our simulations (where there are 20 sensory inputs, each corresponding to a single stimulus dimension), where exact Bayesian inference is likely to be intractable.

Deneve proposed that probabilistic information is encoded through the spike times of individual neurons, rather than their mean firing rates (Deneve, 2008a,b). In common with our work, Deneve hypothesized that the visual system learns a generative model that predicts how binary hidden causes give rise to the received sensory input. While we assumed that individual neurons encode the posterior probability that hidden variables are active ($p(s_i = 1|\boldsymbol{x})$), in Deneve's model, they encode the log-odds ($\log \frac{p(s_i=1|\boldsymbol{x})}{p(s_i=0|\boldsymbol{x})}$). Deneve hypothesized that neurons implement a form of *predictive coding*: each new spike signals an increase in the log-odds that cannot predicted from the neuron's previous spiking output. The neural firing rates predicted by Deneve's model are qualitatively similar to our model: the firing rate of individual neurons correlates with the probability that the encoded hidden variable is active. In common with our work, Deneve et al. predict surround-suppression of visual neuron responses when a stimulus is presented outside of the classical RF (Denéve et al., 2008; Denève and Lochmann, 2009; Lochmann and Deneve, 2011). In the future, a neural implementation of our attentional

model, based on the spiking code proposed by Deneve et al., could be used to investigate the effects of attention on the spiking statistics and temporal dynamics of sensory neurons (see also, section 5.3.3).

5.3.2 Sampling representation of the posterior distribution

Recently, it has been hypothesized that the brain represents the posterior probability distribution using a *sampling code,* where neural responses encode samples from the underlying distribution (Hoyer and Hyvarinen, 2003; Fiser et al., 2010). Thus, rather than being interpreted as 'noise', neural response variability would encode uncertainty in the believed causes of the sensory input. One advantage of this proposal is that, as many classical statistical neural networks use an implicit sampling representation (Hinton and Sejnowski, 1986; Hinton et al., 1995), it is reasonably well understood (although not trivial) how a sampling code could be used to implement learning and representation of a high-dimensional hidden state space (Fiser et al., 2010). However, at present, there has been little work of a more biological bent, showing how probabilistic sampling could be implemented within a realistic neural network (although see (Moreno-Bote et al., 2011; Buesing et al., 2011)).

Several papers have investigated how a sampling code could be distinguished experimentally. At the behavioural level, it has been proposed that perceptual bistability, in which the appearance of a presented stimulus oscillates between different competing interpretations, could come about as a result of probabilistic sampling from the posterior (Schrater and Sundareswara, 2007; Reichert et al., 2011b; Gershman et al., 2009; Moreno-Bote et al., 2011). At the neural level, a sampling representation predicts a close relation between stimulus-evoked and spontaneous neural activity (Fiser et al., 2010). In general, if the internal model is well matched to the external environment, the average posterior distribution should be very similar to the prior, through the identity: $\langle p(s|x)\rangle_{p(x)} = p(s)$. Thus, if evoked and spontaneous neural activities encode samples from the posterior and the prior distributions, respectively, they should also be very similar (Fiser et al., 2010). This prediction has been verified experimentally through electrophysiological recordings of visual cortical activity in awake ferrets (Berkes et al., 2011). Interestingly, the observed similarity between spontaneous and evoked activity was both specific for natural scenes and increased with the animals' age, suggesting that the internal model adapts during development to better reflect the statistics of the environment.

In general, drawing samples from a high-dimensional probability distribution is a challenging problem. In machine learning applications, this is often done using Markov Chain Monte Carlo (MCMC) algorithms, where samples are drawn relative to the current state, and thus, the resulting sequence of samples forms a Markov chain. A limitation of MCMC algorithms is that nearby samples are often highly correlated, so that many samples are needed to represent the posterior. To deal with this limitation, Shelton et al. proposed a biologically plausible

approximate inference algorithm that combines both deterministic and sampling-based approximations (Shelton et al., 2011). Neurally, their algorithm can be interpreted as a feed-forward preselection of the relevant state space (which selects the hidden-states that have a reasonable probability of occurring), followed by a neural implementation of MCMC, to evaluate the posterior over the relevant states. By reducing the size of the hidden-state space, preselection ensures that fewer samples are needed to capture the posterior distribution.

Shelton et al. implemented their algorithm in a binary latent variable model that was very similar to the generative model used in our simulations (section 4.1.2.1). Therefore, a straightforward extension to our work, would be to use the 'select and sample' algorithm proposed by Shelton et al. to construct a biologically plausible sampling implementation of our model. While the predictions for the mean firing rates would be unchanged, the sampling dynamics would make predictions about the variability of neural responses, that are not captured by our present work (Hoyer and Hyvarinen, 2003). Thus, we could investigate how the predicted attention-dependent changes to neural variability and interneuronal correlations compare to what is observed experimentally (Mitchell et al., 2007; Cohen and Maunsell, 2009).

In our work, learning the parameters of the internal model requires evaluating the posterior distribution over the hidden causes, given both the sensory input and reward, $p(s|x,r)$. As pointed out by Sahani (Sahani, 2004), it may be biologically infeasible to evaluate this probability distribution directly (appendix C). However, following Sahani, we show that a sampling approximation, in which samples drawn from the posterior distribution, $p(s|x)$, are multiplied by a reward-dependent weighting factor, can be used to construct a biologically plausible learning algorithm. This algorithm has an interesting neural interpretation. First, visual neurons would encode samples from the posterior distribution, providing input for later brain areas responsible for controlling behaviour. Second, after performing an action, a reward-dependent weighting signal would project back to the sensory cortices, modulating learning. While this learning algorithm is described in fairly abstract terms in our work, a detailed biological implementation, which specified how the reward-dependent feedback is computed and represented in the brain, could be used to make predictions about top-down control of attentional modulation and learning in sensory cortices (Schultz and Dickinson, 2000; Corbetta and Shulman, 2002).

5.3.3 Why the neural code matters for theories of attention

In this thesis, we hypothesized that attention is required due to resource constraints at the computational level, which lead to a mismatch between the agent's internal model and the external environment. In contrast, a neural implementation of our model could be used to investigate how attention is driven by low-level resource constraints, such as the metabolic cost of generating a spike (Laughlin, 2001; Lennie, 2003).

For example, in Deneve's model, a free ('spike-threshold') parameter determines how much

the log-odds for the encoded variable ($\log \frac{p(s_i=1|x)}{p(s_i=0|x)}$) have to increase before a spike is triggered; a smaller threshold produces a more accurate code, but at the cost of an increased number of spikes (Deneve, 2008a). Thus, by dynamically altering the spike-threshold, so that it is reduced for neurons that encode task-relevant stimuli and increased for neurons that encode task-irrelevant stimuli, attention could act to increase the accuracy of the neural representation for task-relevant stimuli while leaving the total number of spikes unchanged[1].

In contrast to our work, where attention-dependent changes in the internal model alter the encoded posterior distribution, here, attention would alter the neural representation of the posterior, but not the posterior distribution itself. These different types of attentional modulation are not mutually exclusive: changes to the internal model could occur alongside changes to the neural code. Indeed, an interesting question is how these different levels interact: how do low-level constraints influence the internal model, and how does the internal model constrain the low-level neural implementation (Gershman and Wilson, 2010)?

5.4 Conclusions

In this thesis, we use a combination of psychophysical experiments and theoretical work to investigate how visual perception and neural responses are influenced by the statistical and behavioural context of presented stimuli. In the literature, contextual changes to visual processing are given a number of different labels (including 'adaptation', 'expectations', 'perceptual learning' or 'attention'), depending on the perceptual or neurophysiological changes that are observed, and the timescale over which they develop. However, the distinction between these cognitive phenomena is often not clear-cut. Further, in trying to understand how they differ, it is easy to get caught in a circular argument. For example, if repulsive perceptual biases are always attributed to sensory 'adaptation' and attractive perceptual biases to 'expectations', then it should be no surprise that adaptation and expectations are found to produce repulsive and attractive perceptual biases, respectively!

In this thesis we treat 'attention' and 'expectations' as *descriptive* terms that refer to the perceptual and neurophysiological *consequences* of varying stimulus statistics and behavioural demands. We define them operationally: changes in perception and neural responses that depend on the presented stimulus statistics are defined as due to a subjects' *expectations*; changes that depend on the subject's behavioural demands are defined as being due to *goal-orientated attention*.

We argue that a Bayesian framework for modelling sensory processing provides a useful language for addressing many questions about how and why visual processing is altered by stimulus context. In this framework, the visual system is assumed to learn an internal model

[1] Of course, there would have to be a corresponding change in the dynamics of upstream neurons that 'read-out' from the neural population (P Seriès and Simoncelli, 2008).

that predicts how explanatory causes in the world generate the received sensory input. We postulate that changes in visual processing associated with expectations and attention reflect optimization of this internal model towards sensory input statistics, and behavioural demands.

An interesting experimental question, is how readily people's prior beliefs about the world adapt in the light of new sensory information. We provide psychophysical evidence indicating that people's prior beliefs are highly adaptable. We find that people quickly adapt their prior expectations following exposure to novel stimulus statistics, and that these learned expectations alter their perception of simple visual features as well as inducing hallucinations when no stimulus is presented.

Another open question is how the internal model is influenced by the behavioural relevance, as well as the statistical context of visual stimuli. This question is crucial for modeling visual attention: attentional modulation of neural responses and perception can occur as a result of changing behavioural demands, in the absence of any changes to the stimulus statistics. We extend previous Bayesian models of visual processing to account for this, hypothesizing that the nervous system learns an internal model that predicts how the sensory input and reward received for performing different actions are generated by a common set of hidden causes. This internal model is assumed to adapt continuously in response to changes in the reward and stimulus statistics. We find that a simple model based on these ideas is able to predict a number of observed effects of attention on visual neuron responses.

Despite the potential of Bayesian models for understanding context-dependent changes to visual processing, a number of basic questions need to be answered before these models represent a truly predictive framework. For example, both the form of the internal model and the neural implementation of Bayesian inference is largely unknown (chapter 5). However, rather than representing a weakness of the modeling framework, the fact that its predictions depend on assumptions about the internal representation and neural code, may ultimately be a strength. It suggests that experimental observations about how perception and neural responses are influenced by stimulus statistics and behavioural demands could be used in the future to constrain Bayesian models of visual processing; thus, helping to answer fundamental questions about about how environmental information is structured and represented in the brain.

Appendix A

Unfolded data

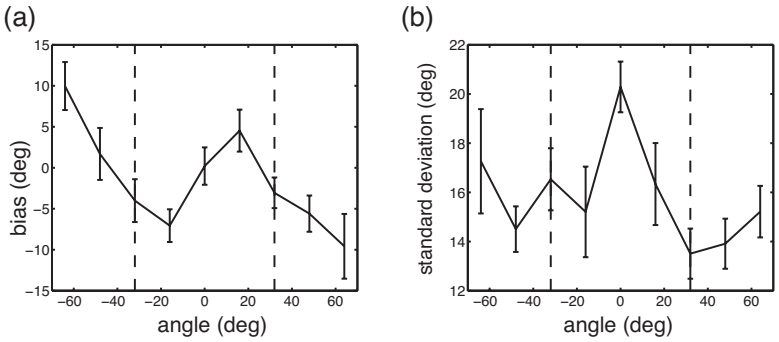


Figure A.1: Average estimation bias (a) and standard deviation of estimation responses (b), plotted against stimulus motion direction. In both plots, results are averaged over all participants and error bars represent within-subject standard error.

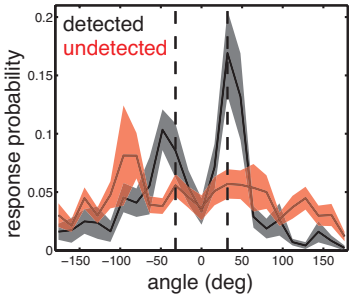


Figure A.2: Probability distributions of participants' estimates of motion direction when no stimulus is present. The two most frequently presented motion directions (±32°) are indicated by vertical dashed lines. Responses were divided into trials where participants reported detecting a stimulus (blue) and trials where they didn't (red). Results are averaged over all participants and error bars represent within-subject standard error.

Appendix B

Gradient of the objective function

To update the parameters of the agent's internal model, we need to compute the gradient of the online objective function:

$$\partial l(\theta,\psi) = \partial \log p(r|a,x,\theta,\psi). \tag{B.1}$$

The derivative of the objective function can be written as:

$$\begin{aligned}\partial l(\theta,\psi) &= \frac{1}{p(r|a,x,\theta,\psi)}\partial p(r|a,x,\theta,\psi)\\ &= \frac{1}{p(r|a,x,\theta,\psi)}\partial \int p(r,s|a,x,\theta,\psi)\,ds.\end{aligned}$$

Taking the derivative inside the integral,

$$\begin{aligned}\partial l(\theta,\psi) &= \frac{1}{p(r|a,x,\theta,\psi)}\int \partial p(r,s|a,x,\psi,\theta)\,ds\\ &= \frac{1}{p(r|a,x,\theta,\psi)}\int p(s|r,a,x,\theta,\psi)\,\partial \log p(r,s|a,x,\psi,\theta)\,ds,\end{aligned}$$

where we have used the identity, $\partial f(x) = f(x)\,\partial \log f(x)$. Rearranging this expression gives,

$$\begin{aligned}\partial l(\theta,\psi) &= \int p(s|r,a,x,\theta,\psi)\,\partial \log p(r,s|a,x,\psi,\theta)\,ds\\ &= \langle \partial \log p(r,s|a,x,\psi,\theta)\rangle_{p(s|x,r,a,\theta,\psi)}\\ &= \langle \partial \log p(r|s,a,x,\psi,\theta)\rangle_{p(s|x,r,a,\theta,\psi)} + \langle \partial \log p(s|x,\theta)\rangle_{p(s|x,r,a,\theta,\psi)}.\end{aligned}$$

The second term in this expression can be expanded as:

$$\begin{aligned}\langle \partial \log p(s|x,\theta)\rangle_{p(s|x,r,a,\theta,\psi)} &= \langle \partial \log p(s,x|\theta)\rangle_{p(s|x,r,a,\theta,\psi)} - \partial \log p(x|\theta)\\ &= \langle \partial \log p(s,x|a,\theta)\rangle_{p(s|x,r,a,\theta,\psi)} - \langle \partial \log p(s,x|\theta)\rangle_{p(s|x,\theta)},\end{aligned}$$

where we have used the identity, $\partial \log p(x|\theta) = \langle \partial \log p(s,x|\theta)\rangle_{p(s|x,\theta)}$. Substituting this back into the expression for the derivative of the objective function, gives:

$$\begin{aligned}\partial l(\theta,\psi) &= \langle \partial \log p(r|s,a,\psi)\rangle_{p(s|x,r,a,\theta,\psi)}\\ &\quad + \langle \partial \log p(s,x|a,\theta)\rangle_{p(s|x,r,a,\theta,\psi)} - \langle \partial \log p(s,x|\theta)\rangle_{p(s|x,\theta)}.\end{aligned}$$

Finally, taking the partial derivative with respect to either θ or ψ returns the expressions shown in the main text:

$$\partial_\psi l(\theta,\psi) = \langle \partial_\psi \log p\,(r|a,s,\psi)\rangle_{p(s|x,r,a,\theta,\psi)} \tag{B.2}$$

$$\partial_\theta l(\theta,\psi) = \langle \partial_\theta \log p\,(s,x|\theta)\rangle_{p(s|x,r,a,\theta,\psi)} - \langle \partial_\theta \log p\,(s,x|\theta)\rangle_{p(s|x,\theta)}. \tag{B.3}$$

Appendix C

Biologically plausible learning algorithm

It is possible that information about both the reward and the sensory input will not be simultaneously available to the organism. In this case they will not be able to compute the posterior distribution of hidden states, conditioned on both the sensory input and the reward ($p(s|x,r,a)$), required to update the model parameters (equations 4.13 and 4.14). This can be dealt with using an importance sampling approximation proposed by Sahani Sahani (2004). Consider the expectation of a function of the hidden states, $g(s)$, over $p(s|r,x,a)$:

$$\langle g(s)\rangle_{p(s|x_n,r_n,a_n)} = \frac{\sum_s g(s)\, p(r_n|s,a_n)\, p(s|x_n)}{\sum_s p(r_n|s,a_n)\, p(s|x_n)}. \tag{C.1}$$

We can approximate this expectation using a sampling algorithm, with N_{samp} samples of s, sampled from the posterior distribution over the hidden states ($s_l \sim p(s|x)$):

$$\begin{aligned}\langle g(s)\rangle_{p(s|x_n,r_n,a_n)} &\approx \frac{\sum_{l=1}^{N_{samp}} g(s_l)\, p(r|s_l,a_n)}{\sum_{l=1}^{N_{samp}} p(r_n|s_l,a_n)} \\ &\approx \sum_{l=1}^{N_{samp}} w_l g(s_l),\end{aligned}$$

where w_l represent the importance weights, $w_l \propto p(r|s_l,a)$ (normalized so that $w_l = 1$). In summary, the expectation over $p(s|x,r,a)$, can be computed by sampling from the posterior distribution $p(s|x)$, and weighting each sample by a factor w_l, depending on the received reward.

We suggest the following neural algorithm for learning the model parameters:

1. The firing rates of visual neurons encode samples from the posterior distribution of hidden causes, given the received sensory input ($s_l \sim p(s|x,\theta)$).

2. These samples are used to estimate the mean reward associated with each action they might perform ($V(a;x,\psi,\theta) \propto \sum_{l=1}^{N_{samp}} \langle r \rangle_{p(r|a,s_l,\psi)}$; possibly encoded in the basal ganglia). The agent performs the action with the highest predicted reward ($\hat{a} = \arg\max_a (V(a;x,\psi,\theta))$).

3. A reward is received, and used to calculate the importance weights for each sample ($w_l \propto p(r|a,s_l,\psi)$). This information is propagated back to earlier sensory areas, to facilitate learning.

4. Synaptic weights (and/or the sensitivity of individual neurons) are updated to follow the gradients, $\sum_{l=1}^{N_{samp}} (w_l - 1) \partial_\theta \log p(s_l,x|\theta)$ and $\sum_{l=1}^{N_{samp}} w_l \partial_\psi \log p(r|s_l,x,a,\psi)$. The impact of each sample on learning is weighted by a reward-dependent factor (w_l), fed-back from areas in the brain which encode reward (e.g. basal ganglia).

Bibliography

Adams, W. J., Graf, E. W., and Ernst, M. O. (2004). Experience can change the 'light-from-above' prior. *Nature Neuroscience*, 7(10):1057–1058.

Anderson, B. (2011). There is no Such Thing as Attention. *Frontiers in Psychology*, 2(September):1–8.

Anderson, B. A., Laurent, P. A., and Yantis, S. (2011). Value-driven attentional capture. *PNAS*, 108(25):10367–71.

Anderson, C. H. and Van Essen, D. C. (1987). Shifter circuits: a computational strategy for dynamic aspects of visual processing. *PNAS*, 84(17):6297–301.

Anstis, S., Verstraten, F. A. J., and Mather, G. (1998). The motion aftereffect. *Trends in Cognitive Sciences*, 2(3):111–117.

Anton-Erxleben, K., Henrich, C., and Treue, S. (2007). Attention changes perceived size of moving visual patterns. *Journal of vision*, 7(11):1–9.

Ball, K. and Sekuler, R. (1982). A specific and enduring improvement in visual motion discrimination. *Science*, 218(4573):697–698.

Bar, M. (2004). Visual objects in context. *Nature Reviews Neuroscience*, 5(8):617–629.

Battaglia, P. W., Di Luca, M., Ernst, M. O., Schrater, P. R., Machulla, T., and Kersten, D. (2010). Within- and cross-modal distance information disambiguate visual size-change perception. *PLoS computational biology*, 6(3):e1000697.

Beck, J. M., Ma, W. J., Kiani, R., Hanks, T., Churchland, A. K., Roitman, J., Shadlen, M. N., Latham, P. E., and Pouget, A. (2008). Probabilistic Population Codes for Bayesian Decision Making. *Neuron*, 60(6):1142–1152.

Berkes, P., Orbán, G., Lengyel, M., and Fiser, J. (2011). Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment. *Science*, 331(6013):83–7.

Berkes, P., Turner, R., and Sahani, M. (2007). On sparsity and overcompleteness in image models. *Advances in Neural Information Processing Systems*, 21.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial vision*, 10(4):433–6.

Broadbent, D. (1958). *Peception and Communication*. Oxford University Press.

Buesing, L., Bill, J., Nessler, B., and Maass, W. (2011). Neural Dynamics as Sampling: A Model for Stochastic Computation in Recurrent Networks of Spiking Neurons. *PLoS Computational Biology*, 7(11):e1002211.

Buiatti, M. and van Vreeswijk, C. (2003). Variance normalisation: a key mechanism for temporal adaptation in natural vision? *Vision Research*, 43(17):1895–1906.

Carandini, M. (2000). Visual cortex: Fatigue and adaptation. *Current biology*, 10(16):R605–7.

Carandini, M., Heeger, D. J., and Movshon, J. A. (1997). Linearity and normalization in simple cells of the macaque primary visual cortex. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 17(21):8621–8644.

Carrasco, M., Ling, S., and Read, S. (2004). Attention alters appearance. *Nature Neuroscience*, 7(3):308–313.

Cave, K. R. and Wolfe, J. M. (1990). Modeling the role of parallel processing in visual search. *Cognitive psychology*, 22(2):225–71.

Chalk, M., Seitz, A. R., and Seriès, P. (2010). Rapidly learned stimulus expectations alter perception of motion. *Journal of vision*, 10(8):2.

Chen, Y., Meng, X., Matthews, N., and Qian, N. (2005). Effects of attention on motion repulsion. *Vision Research*, 45(10):1329–1339.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the acoustical society of America*, 25(5):975–979.

Chikkerur, S., Serre, T., Tan, C., and Poggio, T. (2010). What and where: a Bayesian inference theory of attention. *Vision Research*, 50(22):2233–2247.

Chun, M. M. (2000). Contextual cueing of visual attention. *Trends in Cognitive Sciences*, 4(5):170–178.

Chun, M. M., Golomb, J. D., and Turk-Browne, N. B. (2011). A taxonomy of external and internal attention. *Annual review of psychology*, 62:73–101.

Chun, M. M. and Jiang, Y. (1998). Contextual cueing: implicit learning and memory of visual context guides spatial attention. *Cognitive psychology*, 36(1):28–71.

Chun, M. M. and Nakayama, K. (2000). On the functional role of implicit visual memory for the adaptive deployment of attention across scenes. *Visual Cognition*, 7:65–81.

Clifford, C. W., Wenderoth, P., and Spehar, B. (2000). A functional angle on some after-effects in cortical vision. *Proceedings of The Royal Society of London*, 267(1454):1705–1710.

Cohen, M. R. and Maunsell, J. H. R. (2009). Attention improves performance primarily by reducing interneuronal correlations. *Nature Neuroscience*, 12(12):1594–1600.

Corbetta, M. and Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature reviews. Neuroscience*, 3(3):201–15.

Corteen, R. S. and Dunn, D. (1974). Shock-associated words in a nonattended message: A test for momentary awareness. *Journal of Experimental Psychology*, 102(6):1143–1144.

Courville, A. C., Daw, N. D., and Touretzky, D. S. (2006). Bayesian theories of conditioning in a changing world. *Trends in Cognitive Sciences*, 10(7):294–300.

Crist, R. E., Li, W., and Gilbert, C. D. (2001). Learning to see: experience and attention in primary visual cortex. *Nature Neuroscience*, 4(5):519–525.

Dayan, P. (2008). Load and Attentional Bayes. In *Advances in Neural Information Processing*, pages 369–376.

Dayan, P., Kakade, S., and Montague, P. R. (2000). Learning and selective attention. *Nature Neuroscience*, 3:1218–1223.

Dayan, P. and Solomon, J. A. (2010). Selective Bayes: attentional load and crowding. *Vision research*, 50(22):2248–60.

Dayan, P. and Zemel, R. S. (1999). Statistical models and sensory attention. In *International Conference on Artificial Neural Networks*, volume 2, pages 1017–1022.

Deco, G. (2004). A neurodynamical cortical model of visual attention and invariant object recognition. *Vision Research*, 44(6):621–642.

Deneve, S. (2008a). Bayesian spiking neurons I: inference. *Neural Computation*, 20(1):91–117.

Deneve, S. (2008b). Bayesian spiking neurons II: learning. *Neural Computation*, 20(1):118–145.

Denève, S. and Lochmann, T. (2009). Contextual modulation of visual receptive fields: A Bayesian persperctive. In Trommershauser, J., Körding, K. P., and Landy, M. S., editors, *Sensory cue integration*, pages 448–456.

Denéve, S., Lochmann, T., and Ernst, U. (2008). Spike based inference in a network with divisive inhibition. In *Proc Neurocomp08*.

Desimone, R. (1996). Neural mechanisms for visual memory and their role in attention. *Proceedings of the National Academy of Sciences of the United States of America*, 93(24):13494–9.

Desimone, R. and Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual review of neuroscience*, 18:193–222.

Deutsch, J. A. and Deutsch, D. (1963). Attention: some theoretical considerations. *Psychological Review*, 70(1):80–90.

Downing, C. J. (1988). Expectancy and visual-spatial attention: effects on perceptual quality. *Journal of experimental psychology Human perception and performance*, 14(2):188–202.

Doya, K. (2008). Modulators of decision making. *Nature neuroscience*, 11(4):410–6.

Driver, J. (2001). A selective review of selective attention research from the past century. *British journal of psychology*, 92:53–78.

Duncan, J. (1980). The locus of interference in the perception of simultaneous stimuli. *Psychological Review*, 87(3):273–300.

Duncan, J. (1984). Selective attention and the organization of visual information. *Journal of experimental psychology. General*, 113(4):501–17.

Eckstein, M. P., Abbey, C. K., Pham, B. T., and Shimozaki, S. S. (2004). Perceptual learning through optimization of attentional weighting: Human versus optimal Bayesian learner. *Journal of vision*, 4(12):1006–1019.

Eckstein, M. P., Peterson, M. F., Pham, B. T., and Droll, J. a. (2009). Statistical decision theory to relate neurons to behavior in the study of covert visual attention. *Vision research*, 49(10):1097–128.

Eriksen, B. A. and Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics*, 16(1):143–149.

Ernst, M. O. and Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415(6870):429–433.

Fahle, M. (2005). Perceptual learning: specificity versus generalization. *Current Opinion in Neurobiology*, 15(2):154–160.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27:861–874.

Fischer, B. J. and Peña, J. L. (2011). Owl's behavior and neural representation predicted by Bayesian inference. *Nature neuroscience*, 14(8):1061–1066.

Fiser, J., Berkes, P., Orbán, G., and Lengyel, M. (2010). Statistically optimal perception and learning: from behavior to neural representations. *Trends in Cognitive Sciences*, 14(3):119–130.

Francolini, C. M. and Egeth, H. E. (1980). On the nonautomaticity of "automatic" activation: evidence of selective seeing. *Perception & psychophysics*, 27(4):331–42.

Fuller, S. and Carrasco, M. (2006). Exogenous attention and color perception: performance and appearance of saturation and hue. *Vision research*, 46(23):4032–47.

Ganguli, D. and Simoncelli, E. P. (2010). Implicit encoding of prior probabilities in optimal neural populations. In *Advances in Neural Information Processessing Systems 23*, pages 658–666.

García-Pérez, M. A. (1998). Forced-choice staircases with fixed step sizes: asymptotic and small-sample properties. *Vision Research*, 38(12):1861–1881.

Geisler, W. (2003). A Bayesian approach to the evolution of perceptual and cognitive systems. *Cognitive Science*, 27(3):379–402.

Gekas, N., Chalk, M., and Seriès, P. (2011). Investigating specificity of experimentally induced prior expectations in motion perception. *In preperation*.

Gershman, S. and Wilson, R. C. (2010). The neural costs of optimal control. In *Advances in Neural Information Processessing Systems 23*, pages 712–720.

Gershman, S. J., Cohen, J. D., and Niv, Y. (2010). Learning to Selectively Attend. In *Annual conference of the cognitive science society 32*.

Gershman, S. J. and Niv, Y. (2010). Learning latent structure: carving nature at its joints. *Current Opinion in Neurobiology*, 20(2):251–256.

Gershman, S. J., Vul, E., and Tenenbaum, J. B. (2009). Perceptual multistability as Markov chain Monte Carlo inference. In *Advances in Neural Information Processing 22*, pages 611–619.

Ghose, G. M. (2009). Attentional modulation of visual responses by flexible input gain. *Journal of Neurophysiology*, 101(4):2089–2106.

Ghose, G. M. and Bearl, D. W. (2010). Attention directed by expectations enhances receptive fields in cortical area MT. *Vision Research*, 50(4):441–451.

Ghose, G. M. and Maunsell, J. H. R. (2002). Attentional modulation in visual cortex depends on task timing. *Nature*, 419(6907):616–620.

Girshick, A. R., Landy, M. S., and Simoncelli, E. P. (2011). Cardinal rules: visual orientation perception reflects knowledge of environmental statistics. *Nature neuroscience*, 14:926–932.

Gobell, J. and Carrasco, M. (2005). Attention alters the appearance of spatial frequency and gap size. *Psychological science*, 16(8):644–651.

Gottlieb, J. and Balan, P. (2010). Attention as a decision in information space. *Trends in cognitive sciences*, 14(6):240–8.

Gregory, R. (1970). *The intelligent eye*. London: Weidenfield and Nicolson.

Groos, K. (1896). *Die Spiele der Thiere*.

Grossberg, S. (2000). How hallucinations may arise from brain mechanisms of learning, attention, and volition. *Journal of the International Neuropsychological Society : JINS*, 6(5):583–92.

Haijiang, Q., Saunders, J. A., Stone, R. W., and Backus, B. T. (2006). Demonstration of cue recruitment: change in visual appearance by means of Pavlovian conditioning. *PNAS*, 103(2):483–488.

Hinton, G. E., Dayan, P., Frey, B. J., and Neal, R. M. (1995). The wake-sleep algorithm for unsupervised neural networks. *Science*, 268:1158–1161.

Hinton, G. E. and Sejnowski, T. J. (1986). Learning and relearning in Boltzmann machines. In Rumelhardt, D. E. and McClelland, J. L., editors, *Parallel distributed processing: Explorations in the microstructure of cognition*, pages 282–317. MIT Press.

Hopfinger, J. B., Buonocore, M. H., and Mangun, G. R. (2000). The neural mechanisms of top-down attentional control. *Nature neuroscience*, 3(3):284–91.

Hoyer, P. O. and Hyvarinen, A. (2003). Interpreting neural response variability as Monte Carlo sampling of the posterior. In *Advances in Neural Information Processing Systems 16*, pages 293–300. Citeseer.

Hudson, P. T., van den Herik, H. J., and Postma, E. O. (1997). SCAN: a scalable model of attentional selection. *Neural networks*, 10(6):993–1015.

Hyvärinen, A. (2010). Statistical models of natural images and cortical visual representation. *Topics in Cognitive Science*, 2(2):251–264.

Hyvärinen, A., Hoyer, P. O., Hurri, J., and Gutmann, M. (2005). Statistical models of images and early vision. *Proceedings of the Int. Symposium on Adaptive Knowledge Representation and Reasoning (AKRR2005)*.

James, W. (1890). *The principles of psychology*. Henry Holt, New York.

Jaramillo, S. and Zador, A. M. (2011). The auditory cortex mediates the perceptual effects of acoustic temporal expectation. *Nature neuroscience*, 14(2):246–51.

Jaynes, E. (1986). Bayesian methods: general background. In *Maximum Entropy and Bayesian Methods in Inverse problems*.

Jazayeri, M. (2007). Integration of sensory evidence in motion discrimination. *Journal of vision*, 7(12):1–7.

Jazayeri, M. and Movshon, J. A. (2006). Optimal representation of sensory information by neural populations. *Nature Neuroscience*, 9(5):690–696.

Jazayeri, M. and Movshon, J. A. (2007). A new perceptual illusion reveals mechanisms of sensory decoding. *Nature*, 446(7138):912–915.

Jiang, Y. and Chun, M. M. (2001). Selective attention modulates implicit learning. *The Quarterly Journal of Experimental Psychology*, 54(4):1105–1124.

Johnson, E. N., Hawken, M. J., and Shapley, R. (2008). The orientation selectivity of color-responsive neurons in macaque V1. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 28(32):8096–106.

Kanai, R. and Verstraten, F. a. J. (2005). Perceptual manifestations of fast neural plasticity: motion priming, rapid motion aftereffect and perceptual sensitization. *Vision research*, 45(25-26):3109–16.

Kapadia, M. K., Westheimer, G., and Gilbert, C. D. (2000). Spatial distribution of contextual interactions in primary visual cortex and in visual perception. *Journal of Neurophysiology*, 84(4):2048–2062.

Karklin, Y. and Lewicki, M. S. (2005). A hierarchical Bayesian model for learning nonlinear statistical regularities in nonstationary natural signals. *Neural Computation*, 17(2):397–423.

Karklin, Y. and Lewicki, M. S. (2009). Emergence of complex cell properties by learning to generalize in natural scenes. *Nature*, 457(7225):83–86.

Knill, D. C. (2007). Learning Bayesian priors for depth perception. *Journal of Vision*, 7:1–20.

Knill, D. C. and Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends in Neurosciences*, 27(12):712–719.

Knill, D. C. and Richards, W. (1996a). *Perception as Bayesian Inference*. Cambridge University Press.

Knill, D. C. and Richards, W. (1996b). *Perception as Bayesian inference*.

Komatsu, H. (2006). The neural mechanisms of perceptual filling-in. *Nature Reviews Neuroscience*, 7(3):220–231.

Körding, K. P. and Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature*, 427(6971):244–7.

Körding, K. P. and Wolpert, D. M. (2006). Bayesian decision theory in sensorimotor control. *Trends in cognitive sciences*, 10(7):319–26.

Krauzlis, R. J. and Adler, S. a. (2001). Effects of directional expectations on motion perception and pursuit eye movements. *Visual neuroscience*, 18(3):365–76.

Laughlin, S. B. (2001). Energy as a constraint on the coding and processing of sensory information. *Current opinion in neurobiology*, 11(4):475–80.

Lavie, N. (2005). Distracted and confused?: selective attention under load. *Trends in cognitive sciences*, 9(2):75–82.

Lee, J. and Maunsell, J. H. R. (2009). A normalization model of attentional modulation of single unit responses. *PloS one*, 4(2):e4651.

Lee, T. and Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America A*.

Lennie, P. (2003). The cost of cortical computation. *Current Biology*, 13:493–497.

Levinson, E. and Sekuler, R. (1976). Adaptation alters perceived direction of motion. *Vision Research*, 16(7):779–781.

Liu, Y., Yu, A., and Holmes, P. (2009). Dynamical analysis of Bayesian inference models for the Eriksen task. *Neural Computation*, 21:1520–1553.

Lochmann, T. and Deneve, S. (2011). Neural processing as causal inference. *Current opinion in neurobiology*, 21(5):774–781.

Luck, S. J., Chelazzi, L., Hillyard, S. A., and Desimone, R. (1997). Neural mechanisms of spatial selective attention in areas V1, V2, and V4 of macaque visual cortex. *Journal of Neurophysiology*, 77:24–42.

Lücke, J. and Sahani, M. (2008). Maximal causes for non-linear component extraction. *The Journal of Machine Learning Research*, 9:1227–1267.

Ma, W. J., Beck, J. M., Latham, P. E., and Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience*, 9(11):1432–1438.

Ma, W. J., Beck, J. M., and Pouget, A. (2008). Spiking networks for Bayesian inference and choice. *Current Opinion in Neurobiology*, 18(2):217–222.

Machens, C. K., Gollisch, T., Kolesnikova, O., and Herz, A. V. M. (2005). Testing the efficiency of sensory coding with optimal stimulus ensembles. *Neuron*, 47(3):447–456.

MacKay, D. J. C. (2003). *Information theory, Inference, and Learning Algorithms*. Cambridge University Press.

Marois, R., Leung, H. C., and Gore, J. C. (2000). A stimulus-driven approach to object identity and location processing in the human brain. *Neuron*, 25(3):717–28.

Marr, D. (1982). *Vision; A Computational Investigation into the human representation and processing of visual information*. W. H. Freeman.

Martinez-Trujillo, J. and Treue, S. (2004). Feature-based attention increases the selectivity of population responses in primate visual cortex. *Current Biology*, 14:744–751.

Martinez-Trujillo, J. C. and Treue, S. (2002). Attentional modulation strength in cortical area MT depends on stimulus contrast. *Neuron*, 35(2):365–70.

McAdams, C. J. and Maunsell, J. H. R. (1999). Effects of attention on orientation-tuning functions of single neurons in macaque cortical area V4. *The Journal of neuroscience: the official journal of the Society for Neuroscience*, 19(1):431–441.

McCollough, C. (1965). Color adaptation of edge-detectors in the human visual system. *Science*, 149(3688):1115–1116.

Mishkin, M., Ungerleider, L. G., and Macko, K. A. (1983). Object vision and spatial vision: two cortical pathways. *Trends in Neurosciences*, 6:414–417.

Mitchell, J., Sundberg, K., and Reynolds, J. (2007). Differential attention-dependent response modulation across cell classes in macaque visual area V4. *Neuron*, 55(1):131–141.

Montague, P. R. and King-Casas, B. (2007). Efficient statistics, common currencies and the problem of reward-harvesting. *Trends in Cognitive Sciences*, 11(12):514–519.

Moran, J. and Desimone, R. (1985). Selective attention gates visual processing in the extrastriate cortex. *Science*, 229(4715):782–784.

Moreno-Bote, R., Knill, D. C., and Pouget, A. (2011). Bayesian sampling in visual perception. *PNAS*, 108(30):783–790.

Motter, B. C. (1993). Focal attention produces spatially selective processing in visual cortical areas V1, V2, and V4 in the presence of competing stimuli. *Journal of neurophysiology*, 70(3):909–19.

Mozer, M. C. and Baldwin, D. (2008). Experience-guided search: a theory of attentional control. In *Advances in Neural Information Processessing Systems 20*, pages 1033–1040.

Neisser, U. (1970). *Cognitive psychology*. Appleton-Centure-Crofts, New York.

Newsome, W. T., Britten, K. H., and Movshon, J. A. (1989). Neuronal correlates of a perceptual decision. *Nature*, 341(6237):52–54.

O'Craven, K. M., Downing, P. E., and Kanwisher, N. (1999). fMRI evidence for objects as the units of attentional selection. *Nature*, 401(6753):584–7.

Olshausen, B. a., Anderson, C. H., and Van Essen, D. C. (1993). A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *The Journal of neuroscience*, 13(11):4700–19.

Olshausen, B. A. and Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609.

Olshausen, B. A. and Field, D. J. (2004). Sparse coding of sensory inputs. *Current Opinion in Neurobiology*, 14(4):481–487.

P Seriès, A. A. S. and Simoncelli, E. P. (2008). Is the homunculus 'aware' of sensory adaptation. *Neural Computation*, 21:3271–3304.

Pashler, H. (1998). *The psychology of attention*. MIT Press, Cambridge, Massachusetts.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spatial vision*, 10:437–432.

Pestilli, F. and Carrasco, M. (2005). Attention enhances contrast sensitivity at cued and impairs it at uncued locations. *Vision Research*, 45(14):1867–1875.

Platt, M. L. and Glimcher, P. W. (1999). Neural correlates of decision variables in parietal cortex. *Nature*, 400(6741):233–238.

Polat, U., Mizobe, K., Pettet, M. W., Kasamatsu, T., and Norcia, A. M. (1998). Collinear stimuli regulate visual responses depending on cell's contrast threshold. *Nature*, 391(6667):580–584.

Posner, M. I., Snyder, C. R. R., and Davidson, B. J. (1980). Attention and the detection of signals. *Journal of Experimental Psychology*, 109(2):160–174.

Pouget, A., Zhang, K., Deneve, S., and Latham, P. E. (1998). Statistically efficient estimation using population coding. *Neural Computation*, 10(2):373–401.

Prinzmetal, W., Amiri, H., Allen, K., and Edwards, T. (1998). Phenomenology of attention: 1. color, location, orientation, and spatial frequency. *Perception*, 24(1):261–282.

Prinzmetal, W., Long, V., and Leonhardt, J. (2008). Involuntary attention and brightness contrast. *Perception & Psychophysics*, 70(7):1139–1150.

Prinzmetal, W., Nwachuku, I., Bodanski, L., Blumenfeld, L., and Shimizu, N. (1997). The phenomenology of attention: 2. Brightness and contrast. *Consciousness and cognition*, 6(2-3):372–412.

Puertas, G., Bornschein, J., and Lücke, J. (2010). The maximal causes of natural scenes are edge filters. In *Advances in Neural Information Processing 23*, pages 1939–1947.

Rangel, A., Camerer, C., and Montague, P. R. (2008). A framework for studying the neurobiology of value-based decision making. *Nature Reviews Neuroscience*, 9(7):545–556.

Rao, R. P. and Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1):79–87.

Rao, R. P. N. (2004). Bayesian computation in recurrent neural circuits. *Neural Computation*, 16(1):1–38.

Rao, R. P. N. (2005). Bayesian inference and attentional modulation in the visual cortex. *NeuroReport*, 16(16):1843–1848.

Rao, R. P. N., Olshausen, B. A., and Lewicki, M. S. (2002). *Probabilistic models of the brain: perception and neural function*. MIT Press.

Rauber, H. J. and Treue, S. (1998). Reference repulsion when judging the direction of visual motion. *Perception*, 27(4):393–402.

Raymond, J. (2000). Attentional modulation of visual motion perception. *Trends in Cognitive Sciences*, 4(2):42–50.

Reichert, D. P., Series, P., and Storkey, A. (2011a). A hierarchical generative model of recurrent object-based attention in the visual cortex. In *ICANN*, pages 18–25.

Reichert, D. P., Seriès, P., and Storkey, A. J. (2011b). Neuronal adaptation for sampling-based probabilistic inference in perceptual bistability. In *Advances in Neural Information Processing Systems 24*, pages 2357–2365.

Reynolds, J. H. and Chelazzi, L. (2004). Attentional modulation of visual processing. *Annual Review of Neuroscience*, 27:611–647.

Reynolds, J. H., Chelazzi, L., and Desimone, R. (1999). Competitive mechanisms subserve attention in macaque areas V2 and V4. *The Journal of neuroscience*, 19(5):1736–1753.

Reynolds, J. H. and Desimone, R. (1999). The role of neural mechanisms of attention in solving the binding problem. *Neuron*, 24(1):19–29, 111–25.

Reynolds, J. H. and Heeger, D. J. (2009). The normalization model of attention. *Neuron*, 61(2):168–185.

Reynolds, J. H., Pasternak, T., and Desimone, R. (2000). Attention increases sensitivity of V4 neurons. *Neuron*, 26(3):703–714.

Roberts, M., Delicato, L. S., Herrero, J., Gieselmann, M. A., and Thiele, A. (2007). Attention alters spatial integration in macaque V1 in an eccentricity-dependent manner. *Nature Neuroscience*, 10(11):1483–1491.

Roberts, M. J. and Thiele, A. (2008). Attention and contrast differently affect contextual integration in an orientation discrimination task. *Experimental brain research*, 187(4):535–549.

Roelfsema, P. R., Lamme, V. A., and Spekreijse, H. (1998). Object-based attention in the primary visual cortex of the macaque monkey. *Nature*, 395(6700):376–381.

Sahani, M. (2004). A biologically plausible algorithm for reinforcement-shaped representational learning. In *Advances in Neural Information Processing 16*, pages 1287–1294.

Sahani, M. and Dayan, P. (2003). Doubly distributional population codes: simultaneous representation of uncertainty and multiplicity. *Neural Computation*, 15(10):2255–2279.

Sahani, M. and Whiteley, L. (2007). A unifying probabilistic computational framework for attention. In *Cosyne*.

Sahani, M. and Whiteley, L. (2011). Modeling cue integration in cluttered environments. In *Sensory cue integration*. Oxford University Press.

Schneider, K. A. (2006). Does attention alter appearance? *Perception & Psychophysics*, 68(5):800–814.

Schneider, K. A. and Komlos, M. (2008). Attention biases decisions but does not alter appearance. *Journal of vision*, 8(15):1–10.

Schrater, P. R. and Sundareswara, R. (2007). Theory and dynamics of perceptual bistability. In *Advances in Neural Information Processing Systems 19*, volume 19, pages 1217–1224. Citeseer.

Schultz, W. (2006). Behavioral theories and the neurophysiology of reward. *Annual review of psychology*, 57:87–115.

Schultz, W., Dayan, P., and Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, 275(5306):1593–1599.

Schultz, W. and Dickinson, A. (2000). Neuronal coding of prediction errors. *Annual review of neuroscience*, 23:473–500.

Schwartz, O., Hsu, A., and Dayan, P. (2007). Space and time in visual context. *Nature Reviews Neuroscience*, 8(7):522–535.

Schwartz, O., Sejnowski, T. J., and Dayan, P. (2006). Soft mixer assignment in a hierarchical generative model of natural scene statistics. *Neural Computation*, 18(11):2680–2718.

Schwartz, O., Sejnowski, T. J., and Dayan, P. (2009). Perceptual organization in the tilt illusion. *Journal of vision*, 9(4):19.1–20.

Schwartz, O. and Simoncelli, E. P. (2001). Natural signal statistics and sensory gain control. *Nature Neuroscience*, 4(8):819–825.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Scolari, M. and Serences, J. T. (2009). Adaptive allocation of attentional gain. *The Journal of neuroscience*, 29(38):11933–42.

Seitz, A., Yamagishi, N., Werner, B., Goda, N., and Kawato, M. (2005a). Task-specific disruption of perceptual learning. *PNAS*, 102(41):14895–14900.

Seitz, A. R., Kim, D., and Watanabe, T. (2009). Rewards evoke learning of unconsciously processed visual stimuli in adult humans. *Neuron*, 61(5):700–707.

Seitz, A. R., Nanez, J. E., Holloway, S. R., Koyama, S., and Watanabe, T. (2005b). Seeing what is not there shows the costs of perceptual learning. *PNAS*, 102(25):9080–9085.

Sekuler, R. and Ball, K. (1977). Mental set alters visibility of moving targets. *Science (New York, NY)*, 198(4312):60–62.

Seriès, P., Lorenceau, J., and Frégnac, Y. (2003). The silent surround of V1 receptive fields: theory and experiments. *Journal of Physiology*, 97(4-6):453–474.

Seung, H. S. and Sompolinsky, H. (1993). Simple models for reading neuronal population codes. *PNAS*, 90(22):10749–10753.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27:623–656.

Shelton, J. A., Bornschein, J., Sheikh, A. S., Berkes, P., and Lücke, J. (2011). Select and sample - a model of efficient neural inference and learning. In *Advances in Neural Information Processessing Systems 24*.

Shi, L. and Griffiths, T. L. (2009). Neural Implementation of Hierarchical Bayesian Inference by Importance Sampling. In *Advances in Neural Information Processessing Systems 22*, pages 1669–1677.

Simoncelli, E. (2009). Optimal estimation in sensory systems. In *The New Cognitive Neurosciences*, chapter 36. MIT Press.

Simoncelli, E. P. and Heeger, D. J. (1998). A model of neuronal responses in visual area MT. *Vision Research*, 38(5):743–761.

Simoncelli, E. P. and Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience*, 24:1193–1216.

Simoncelli, E. P. and Schwartz, O. (1999). Modeling surround suppression in V1 neurons with a statistically derived normalization model. *Advances in Neural Information Processing Systems*, pages 153–159.

Sotiropoulos, G., Seitz, A. R., and Seriès, P. (2011). Changing expectations about speed alters perceived motion direction. *Current Biology*, 21(21):R884.

Spitzer, H., Desimone, R., and Moran, J. (1988). Increased attention enhances both behavioral and neuronal performance. *Science (New York, NY)*, 240(4850):338–340.

Sterzer, P., Frith, C., and Petrovic, P. (2008). Believing is seeing: expectations alter visual awareness. *Current biology*, 18(16):R697–8.

Stocker, A., Simoncelli, E., Platt, J., and Koller, D. (2006). A Bayesian Model of Conditioned Perception. In *Advances in Neural Information Processessing Systems 20*, pages 1409–1416.

Stocker, A. A. and Simoncelli, E. P. (2006a). Noise characteristics and prior expectations in human visual speed perception. *Nature Neuroscience*, 9(4):578–585.

Stocker, A. A. and Simoncelli, E. P. (2006b). Sensory adaptation within a Bayesian framework for perception. In *Advances in Neural Information Processing 18*, pages 1291–1298.

Summerfield, C. and Egner, T. (2009). Expectation (and attention) in visual cognition. *Trends in Cognitive Sciences*, 13(9):403–409.

Sun, J. and Perona, P. (1998). Where is the sun? *Nature neuroscience*, 1(3):183–4.

Sundberg, K. A., Mitchell, J. F., and Reynolds, J. H. (2009). Spatial attention modulates center-surround interactions in macaque visual area V4. *Neuron*, 61(6):952–963.

Sutherland, S. (1998). Feature selection. *Nature*, 25(3):359.

Sutton, R. S. and Barto, A. G. (1998). *Reinforcement learning: an introduction*. MIT Press.

Thiele, A., Dobkins, K. R., Albright, T. D., and Jolla, L. (2001). Neural Correlates of Chromatic Motion Perception Salk Institute for Biological Studies. *Neuron*, 32:351–358.

Treisman, A. M. (1960). Contextual cues in selective listening. *The Quarterly Journal of Experimental Psychology*, 12:242–248.

Treisman, A. M. (1969). Strategies and models of selective attention. *Psychological Review*, 73(6):282–299.

Treisman, A. M. and Gelade, G. (1980). A feature-integration of attention. *Cognitive Psychology*, 12:97–136.

Treue, S. and Martínez Trujillo, J. C. (1999). Feature-based attention influences motion processing gain in macaque visual cortex. *Nature*, 399(6736):575–579.

Tsotsos, J. K. (1989). The Complexity of Perceptual Search Tasks. In *Proceedings, International Joint Conference on Artificial Intellegence*, pages 1571–1577.

Tsotsos, J. K. (1990). Analyzing vision at the complexity level. *Behavioral and Brain Sciences*, 13:423–469.

Tsotsos, J. K., Culhane, S. M., Kei, W. Y., Lai, Y., Davis, N., and Nuflo, F. (1995). Modeling visual attention via selective tuning. *Artificial Intelligence*, 78:507–545.

Turatto, M., Vescovi, M., and Valsecchi, M. (2007). Attention makes moving objects be perceived to move faster. *Vision Research*, 47(2):166–178.

Tzvetanov, T., Womelsdorf, T., Niebergall, R., and Treue, S. (2006). Feature-based attention influences contextual interactions during motion repulsion. *Vision Research*, 46(21):3651–3658.

Wainwright, M. J., Schwartz, O., and Simoncelli, E. P. (2001). Natural image statistics and divisive normalization: modeling nonlinearity and adaptation in cortical neurons. In *Probabilistic models of the brain: perception and neural function*, pages 203–222. MIT Press.

Weiss, Y., Simoncelli, E. P., and Adelson, E. H. (2002). Motion illusions as optimal percepts. *Nature Neuroscience*, 5(6):598–604.

Whiteley, L. (2008). *Uncertainty, reward, and attention in the Bayesian brain*. PhD thesis, University College London.

Williford, T. and Maunsell, J. H. R. (2006). Effects of spatial attention on contrast response functions in macaque area V4. *Journal of Neurophysiology*, 96(1):40–54.

Wolfe, J. M., Cave, K. R., and Franzel, S. L. (1989). Guided search: an alternative to the feature integration model for visual search. *Journal of experimental psychology. Human perception and performance*, 15(3):419–33.

Yantis, S. (2000). Goal-directed and stimulus-driven determinants of attentional control. In *Control of cognitive processes: attention and performance XVIII*, pages 73–103.

Yantis, S. and Serences, J. T. (2003). Cortical mechanisms of space-based and object-based attentional control. *Current Opinion in Neurobiology*, 13(2):187–193.

Yoshiura, T., Zhong, J., Shibata, D. K., Kwok, W. E., Shrier, D. a., and Numaguchi, Y. (1999). Functional MRI study of auditory and visual oddball tasks. *Neuroreport*, 10(8):1683–8.

Yu, A. and Dayan, P. (2005a). Inference, attention, and decision in a Bayesian neural architecture. In *Advances in Neural Information Processing Systems 17*, pages 1577–1584.

Yu, A. and Dayan, P. (2005b). Uncertainty, Neuromodulation, and Attention. *Neuron*, 46(4):681–692.

Yu, A. J., Dayan, P., and Cohen, J. D. (2009). Dynamics of attentional selection under conflict: toward a rational Bayesian account. *Journal of experimental psychology: human perception and performance*, 35(3):700–17.

Yuille, A. L. and Bulthoff, H. H. (1996). Bayesian decision theory and Psychophysics. In Knill, D. C., editor, *Perception as Bayesian inference*, chapter 5. Cambridge University Press.

Zemel, R. S., Dayan, P., and Pouget, a. (1998). Probabilistic interpretation of population codes. *Neural computation*, 10(2):403–30.

Zhang, P., Bao, M., Kwon, M., He, S., and Engel, S. A. (2009). Effects of orientation-specific visual deprivation induced with altered reality. *Current biology*, 19(22):1956–1960.

Zhaoping, L. and May, K. a. (2007). Psychophysical tests of the hypothesis of a bottom-up saliency map in primary visual cortex. *PLoS computational biology*, 3(4):e62.